UNIVERSITY OF SZEGED

DOCTORAL SCHOOL OF EDUCATION

LEARNING AND INSTRUCTION

ASSESSING STUDENTS' SCIENCE MISCONCEPTIONS AND INDUCTIVE REASONING: CROSS-SECTIONAL STUDIES IN INDONESIAN CONTEXT

PhD Dissertation

Soeharto Soeharto

Supervisor:

Prof. Dr. Benő Csapó

Professor of Education

SZEGED, HUNGARY

2023

TABLE OF CONTENTS

ABBREVATIONS

OECD	The Organization for Economic Co-operation and Development
PISA	Programme for International Student Assessment
PSTs	preservice science teachers
NA	Number analogies
NS	Number series
FS	Figural series
FA	Figural analogies
CFA	Confirmatory factor analysis
EFA	Exploratory factor analysis
RMSEA	Root-Mean-Square Error of Approximation
SPSS	Statistical Package for Social Sciences
SRMR	Standardized Root Mean Square Residual
CFI	Comparative fit index
ANOVA	Analysis of Variance
MOEC	Ministry of Education and Culture
PKMAPs	Person diagnostic maps
ULS	Unweighted least squares
MNSQ	Mean squares
PTMA	Point measure correlation
ZSTD	Z-standardized
DIF	Differential item functioning

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 Background and context of the study

In the PISA report 2018 (OECD, 2020) for student performances in science, Indonesian students performed the worst out of 79 nations, which may indicate that most Indonesian students struggle to understand scientific notions during the learning process. Numerous studies (e.g., (Arslan et al., 2012; Keeley, 2012; Mubarokah et al., 2018; Samsudin et al., 2021) have demonstrated the connection between scientific misconceptions and student academic achievement and how they affect student learning activity in science Hence, it stands to reason that if students struggle to understand a particular scientific subject, they will likely struggle in the future or during the learning process, which will lead to poor science achievement.

Through classroom instruction and outside learning, students build their knowledge. Students have prior knowledge, abilities, and experience that shape their initial notions in scientific learning before engaging in a learning activity at school. Although this condition still exists after the science learning activity is completed, these initial conceptions may be in conflict with scientific concepts, called misconceptions in science (Eshach et al., 2018; Köse, 2004; Stefanidou et al., 2019). Misconceptions are ideas that contradict or do not adhere to scientific principles (Martin, 2005). According to Allen (2014), a person's knowledge that is based on formal and informal experiences unrelated to scientific knowledge is a misconception. In addition, as science and technology advance quickly, the depth of information grows, changing the meaning of scientific notions (Arslan et al., 2012; Kiray & Simsek, 2021). Students' conceptual understanding is impacted by incorrect and incomplete knowledge brought about by student experience, misinformation in teacher learning, and misunderstandings in the analysis of information in textbooks (Kirbulut & Geban, 2014; Zlatkin-Troitschanskaia et al., 2015).

Various studies have been conducted to determine the different science-learning ideas that cause student misconceptions in science. Wandersee et al. (1994) analyzed 103 studies related to misconceptions, and Gurel et al. (2015) discovered 273 articles about misconceptions. There are three publications (Fajarini et al., 2018; Fariyani et al., 2017; Ratnasari & Suparmi, 2017) that talk about detecting student misconceptions in Indonesia and how this relates to the dearth of research issues in the country's field of scientific education. However, these recent Indonesian publications mainly focused on identifying student misconceptions in a single science concept, such as global warming, optics, or heat, and there is no instrument now being developed from science ideas dispersing student misunderstanding in learning science. There is also lack of evidence how is the pattern of misconceptions in sciences and how are students' ability in solving science concepts. Therefore, there is a need to investigate Indonesian students' misconceptions in science with various background factors such as gender and grade levels.

In addition of student misconceptions in science in this present research, there is an interest area than can be explored related to student ability in science in classroom contexts such student thinking skills, especially inductive reasoning. Inductive reasoning generally correlates with mathematics, reading, and science (De Koning et al., 2002; Nikolov & Csapó, 2018; Van Vo & Csapó, 2022). Indonesia's low ranking in science domain in the 2018 PISA report (OECD, 2020), 71st out of 79 participating nations, may be the results of students' low inductive reasoning ability.

In Indonesia, the 2013 Indonesian core curriculum included thinking skills (Hasan, 2013; Prastowo & Fitriyaningsih, 2020). The learning material was created to link to the fundamental competencies in several disciplines in the three primary domains of attitude, skills, and knowledge supported by this curriculum (Hasan, 2013). This curriculum has a significant issue with evaluation practices, particularly when assessing attitude. It was challenging to adjust the attitude assessment to the setting of the classroom because it was brand-new. According to Badaruddin & Hawi (2022), the majority of teachers expressed frustration over how challenging it was to gauge student attitudes and that their knowledge of the best methods and evaluation tools was still lacking. However, it was simple to evaluate knowledge and abilities (Natsir et al., 2018). The teacher may use a variety of learning models on various resources and subjects to improve students' thinking skills (Prastowo & Fitriyaningsih, 2020). The inductive reasoning test has been used in the general basic skills knowledge test when applying for positions at the government and corporate levels, despite the fact that it is not taught and trained explicitly in schools. Limited data and studies were related to inductive reasoning in classrooms and even in institutes of higher learning. Therefore, there is a need to perform an evaluation of Indonesian student inductive reasoning to be a pioneer for assessment of inductive reasoning in Indonesia.

In Indonesian curriculum development framework as presented in Figure 1, there are also no specific details in nurturing student science misconception and inductive reasoning skills. Psychology, Pedagogy and Socio-eco-cultural became the main target in the curriculum development framework. The pedagogical part only focus on feasibility including teaching material, teaching method, and assessment. Therefore, the context in this study can give additional values and information to cover students' misconceptions and inductive reasoning skills that not embedded in the Indonesian curriculum development framework. Consequently, evaluation of student misconceptions in science and inductive reasoning skills in Indonesian context are topics that can offer initial information and a foundation for further studies in educational area.

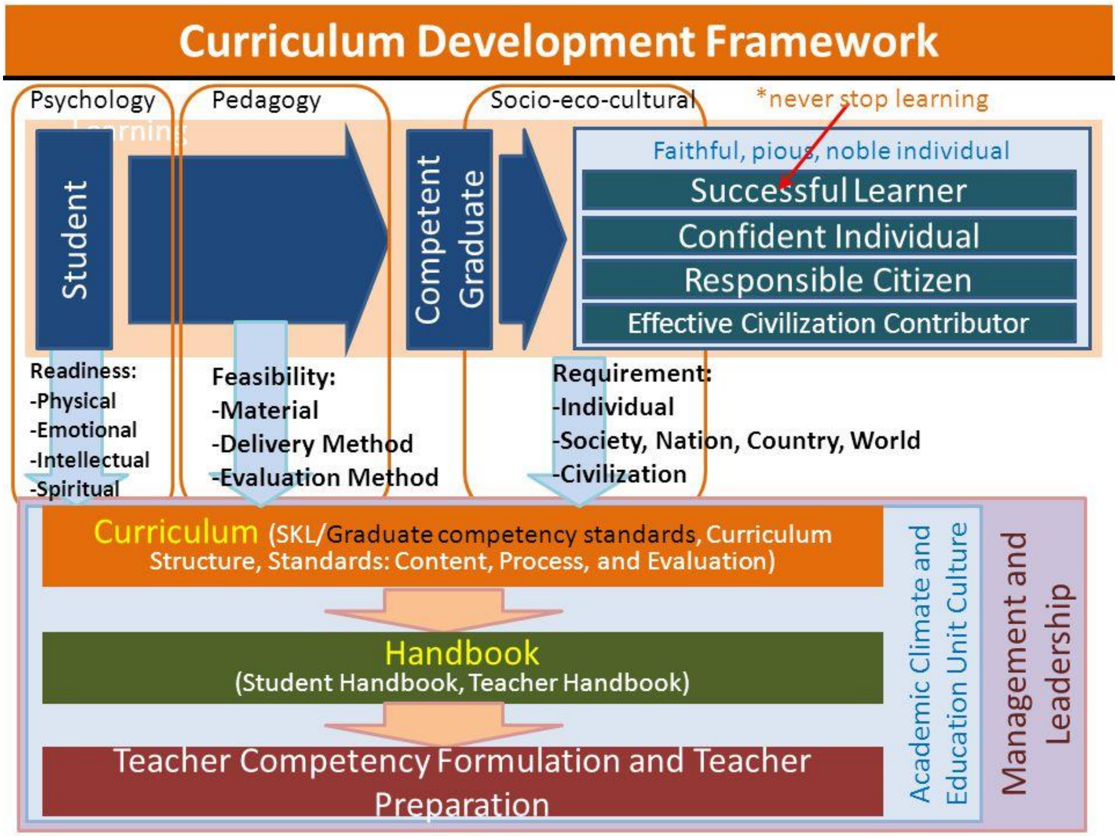

Curriculum Development Framework
Psychology
Pedagogy
Socio-eco-cultural
*never stop learning
Student
Competent Graduate
Faithful, pious, noble individual
Successful Learner
Confident Individual
Responsible Citizen
Effective Civilization Contributor
Readiness:
-Physical
-Emotional
-Intellectual
-Spiritual
Feasibility:
-Material
-Delivery Method
-Evaluation Method
Requirement:
-Individual
-Society, Nation, Country, World
-Civilization
Curriculum (SKL/Graduate competency standards, Curriculum Structure, Standards: Content, Process, and Evaluation)
Handbook
(Student Handbook, Teacher Handbook)
Teacher Competency Formulation and Teacher Preparation
Academic Climate and Education Unit Culture
Management and Leadership

Figure 1. The Indonesian curriculum development framework.

1.2 Statement of the problem

Indonesia implemented the 2013 curriculum for more than 10 years. This curriculum focuses on three domains namely attitude, skills, and knowledge. However, there is no specific assessment to identify students' misconceptions in science and inductive reasoning skills. Whereas both constructs are important in guiding students' achievement in academic and work field. To start the investigation of students' ability to understand science concepts and inductive reasoning skills before investigating the structural model or causal relationship stage. There is a need of assessment in comprehensive work to pioneer this research topic.

In addition, the literature review conducted by Soeharto et al. (2019) have confirmed that topics of physics, chemistry, and biology subject in science were distributing misconception for the student in Indonesia from 111 published studied reviewed. However, only four studies that measure misconceptions in science. The studies of inductive reasoning in Indonesian context are also limited in schools and higher education context (Istikomah et al., 2017; Siswono et al., 2020). Furthermore, the inductive reasoning test has been used in the general basic skills knowledge test when applying for jobs at the government and company levels, even though inductive reasoning is not explicitly taught and studied in schools. Therefore, there is a need to do assessment to identify student misconceptions in science and inductive reasoning skills in Indonesian context.

The research project in this dissertation consists two cross-sectional studies from pilot and main study. One systematic literature review and four empirical stduies had been published as dissertation output. Firstly, a systematic review was conducted to identify what kinds of topics in science concepts that possible to distributing misconception. Secondly, the pilot study was used to develop instruments in two-tier multiple choice to measure student misconception in sciences. Then, a specific study was conducted to investigate the item difficulties in science and student answer pattern. The fourth study, the exploration of student misconception in science was conducted. The last study, the assessment of student inductive reasoning was also performed to pioneer and inform researchers, educators, and stakeholders about Indonesian student inductive reasoning skills.

1.3 Organization of the dissertation

This dissertation is composed based on two cross sectional studies from pilot and main study with five different published studies in the assessment topic of student misconceptions in science and inductive reasoning skills (Soeharto, 2021; Soeharto et al., 2019; Soeharto & Csapó, 2021, 2022b, 2022a). The dissertation consists of five different chapters. Chapter one is the introduction which consists of the study context, statement of the problem and organization of dissertation. Chapter two is a review of literature on studies related to the research topic in this dissertation. The focus was on assessment of student misconception in sciences and inductive reasoning skills in

Indonesian context. Chapter three focuses on study aims, research question, structure of empirical studies, and the methodology section was used in the empirical studies which focus on design, sampling procedure, data collection, data analysis, instrument and validation. Chapter four presents four empirical studies in this dissertation. The systematic review of students' common misconceptions in science and its' diagnostic assessment tools was included in chapter two. This systematic literature review focuses on initial investigation of topics in science causing student misconception and what kind of instruments used in previous studies. The first empirical study is the evaluation and development of students' misconception using diagnostic assessment in science across school grades. This study actually a pilot study as an initial stage in developing two-tier multiple choice test in measuring student misconceptions in science. Study two is the evaluating item difficulty patterns for assessing student misconceptions in science across physics, chemistry, and biology concepts. This study focuses in investigating item difficulty patterns across the science subject using Rasch measurement approach. Study three is an investigation of Indonesian student misconceptions in science concepts in specific using Rasch measurement approach. The last study, study four is a comprehensive assessment of Indonesian inductive reasoning skills and validation of inductive reasoning test using Rasch measurement approach. All five empirical studies have been published in Scopus journals in Q1 and Q2 tier with double-blind peer review system. Chapter five is the conclusion, educational implication, recommendations, and limitations.

CHAPTER 2. LITERATURE REVIEW

2.1 General overview

This chapter has ten different sections that cover both misconceptions in science and inductive reasoning to emit reader understanding in the dissertation. Section 2.1. provide a general overview of all other sections in Chapter 2 to provide a more detailed overview. Section 2.1 gives definition and research applications about misconceptions in science. This section is useful to understand the topic and previous research practices related to misconceptions in various science disciplines. Section 2.3 is the systematic review related to common science concepts that distribute misconceptions in science learning and review what is the trend of instrument used to assess misconception in science. Section 2.4 and Section 2.5 focus on giving an overview based on target participant students and pre-service teacher (University student) that usually carry misconception in understanding science concepts. Section 2.6 and Section 2.7 focus on the developed instrument in the dissertation. A two-tier multiple choice diagnostic test was developed as main research instrument in this dissertation. Section 2.8 provides an overview of inductive reasoning regarding definition and role in education area. Section 2.9 explains the role of inductive reasoning in the student development that covers both theory and practice where students as participants. Section 2.10 highlights the research practices in assessing inductive reasoning in various contexts starting from meta-analysis, empirical research, and cross-cultural adaptation. Finally, Section 2.11 describe the possibility of the relationship between students' misconceptions, science achievement, and inductive reasoning. The role of analogic and analogical reasoning in connection to misconception are also covered in this section.

2.2 Misconceptions in science

Concepts are ideas formed objects or abstraction helping the individual to comprehend the scientific world phenomena (Eggen et al., 2007). Misconceptions are delineated as ideas or insights from students who provide incorrect meaning constructed based on an event or person experience (Martin, 2005). Science misconception is an individual knowledge which is gained from educational experience or informal event that is not relevant or not having the meaning according to scientific concepts (Allen, 2014). In summary, misconceptions in science can be described as student ideas from life experience or informal education which is not structured well resulting in the incorrect meaning according to a scientific concept.

A misconception is categorized into five types namely preconceived notions, non-scientific beliefs of conceptual misunderstandings, conceptual misunderstandings, vernacular misconceptions, and factual misconceptions (Keeley, 2012; Kerr et al., 2006; Leaper et al., 2012; Murdoch, 2018). Preconceived notions are popular conceptions that come from life and personal experience (Murdoch, 2018), for example, many people believing that to see an object, light must first hit our eyes even though the opposite. Preconceived notions occur because students have not yet learned the concept of light. Non-scientific beliefs are views or knowledge acquired by students other than scientific sources (Kerr et al., 2006), for example, some people believe that gender differences determine the ability of students to learn mathematics, science, and language so that men become dominant compared to women. Conceptual misunderstandings are scientific information that arises when students construct their own confusing and wrong ideas based on the correct scientific concepts (Allen, 2014; Keeley, 2012), for example, students find it difficult to understand the concept of normal style because they only understand that style is only a push and a pull. Vernacular misconceptions are mistakes that arise from the use of words in everyday life that have different meanings based on scientific knowledge (Keeley, 2012), for example, students have difficulty to comprehend the concept of heat because students do not understand that heat arises due to regulating rising not only because of fire. Factual misconceptions are misunderstandings that occur at an early age and maintained until adulthood. For instance, children believe they will be struck by lightning if they are outside the house (Eshach et al., 2018).

There are various studies related to students' misconception on learning science because misconception in science concept will be barrier to learn science in current and further level of learning activity. Students' misconceptions in science are persistent, resistant to change and deeply rooted in some science concepts (Wandersee et al., 1994). If students get misconception in science concept, they will be difficult to understand some concepts in current study or future study, and misconception leads student to get low academic achievement.

In science, if students' misconceptions were identified properly, the teacher can help students to know better knowledge understanding on scientific conceptions. Helping students to revise their conception and develop meaningful scientific understanding will make science more useful in their everyday lives about how important is learning science in school. Because of the essential role to revise, reduce and identify students’ misconceptions in science, this study gives some knowledge to help researcher know which concepts in science usually distribute misconception to students in order to overcome students’ misconceptions and to know diagnostic assessment tools that are commonly used to identify misconception in science (Csapó, 2012; Csapó & Szabó, 2012; Soeharto et al., 2019).

Educators and scholars have described different conceptual changes experienced by an individual derived from their intuitive beliefs, life experiences, cultural influences and learning and teaching processes (Arslan et al., 2012; Galvin & Mooney, 2015; Keeley, 2012). Different terminologies and meanings regarding the nature of students’ conceptual understanding reflect the application of misconceptions in various research areas (Kaltakci-Gurel et al., 2017). Misconceptions in science learning have been constantly studied because such misconceptions are persistent, resistant to change in students’ minds and rooted in science concepts (Taslidere, 2016; Treagust, 1988; Treagust & Duit, 2008; Wandersee et al., 1994). In addition, if students experience misconceptions or fail to correctly understand science concepts, they would find it difficult to understand and solve science problems, which would lead to low academic attainment in science disciplines (Mintzes et al., 2005). Students’ misconceptions connected to science concepts need to be identified early so that teachers can construct knowledge that meets competency requirements in science learning. Hence, misconceptions in science concepts are pivotal and essential to investigate in the science education field.

2.3 A review of students’ common misconceptions in science and its’ diagnostic assessment tools: systematic literature review

2.3.1 Introduction

Students learn the concept of knowledge about the world around them form an education system at schools or informal way according to their experience. The experience from students is frequently used to construct an insight with student perspectives. Because of that matter, some studies had been held to provide information about student understanding, especially in learning science concepts. The different insight of student concepts had been defined by a number of terms like “alternative conceptions” (Wandersee et al., 1994), “conceptual difficulties” (Stefanidou et al., 2019), “misconceptions” (Eshach et al., 2018), “mental models” (Wuellner et al., 2017), and others. This study is a literature review of assessment and misconceptions in learning science. It will give information about a) definition of science misconception, b) role of misconception in the instruction process, and c) type of misconceptions in science education.

National Research Council (1997) states that the primary role of misconceptions in science is a barrier for students to learning science because in many cases, misconception can detain students to develop correct ideas used as the initial insight for advance learning. This argument is supported with other findings, King (2010) found that misconception found in textbook of Earth Science influencing students to understand a text in science which make student difficult to understanding further information or knowledge as a reader. A misconception is also affecting teacher understanding in science, so some teachers who have misconception giving implication to practice in teaching (Moodley & Gaigher, 2019), the same case also found in teachers who were teaching physics, chemistry, and biology topics (Bektas, 2017). In a simple explanation, we can say that misconception will interfere with the quality and quantity learning process and outcomes in science for student and teacher.

This study aims to review concepts that often distribute misconceptions to students and diagnostic instruments commonly used to research from 2015 to 2019. With knowing the concept of science accurately distributed misconceptions to student and diagnostic instruments used to, researchers will be easier to conduct research and improve the quality of research.

Research related to students' misconceptions in science is essential because of the following reasons. (1) Research on students' misconceptions on science education had become popular topics over the last four decades. (2) Students' misconceptions are wrong or false concepts had by students. It becomes prominent, reliable and persistent in every topic in science. The best way to improve student understanding in science is to deal with their misconceptions at the first step. (3) Diagnostic instruments or tests are assessment tools concerned with identifying students' misconception in science. The tests are available on many forms such as interview, multiple-choice question, open-ended question, multi-tier question, and others. It will be easy to conduct research if researcher know the benefits and drawback for each instrument.

This study has three mains objective namely, (1) to find common misconception topics distributed misconceptions to student, (2) to analyze diagnostic instrument used to identify students' misconception in science education, (3) the benefits and drawbacks of all diagnostic instruments in research conducted by researcher before. This study also offers some contributions for the future research namely, (1) providing an overview of the scientific topic in learning that are naturally studied and provide misconceptions to student, (2) giving summary for all diagnostic instruments according their benefits and drawbacks in assessing misconception in science, (3) presenting quantitative data for which instrument used to identify student misconception in science education.

2.3.2 Method of systematic literature review

A systematic and structured literature review was used to analyze, examine and describe the current empirical studies on students' misconceptions in science education. To confirm that process review is systematic, we had been using the Preferred Items for Systematic Reviews and Meta Analysis (PRISMA) statement (Moher et al., 2009), with the following steps: (1) establishing criteria for the subject and defining relevant studies; (2) searching strategy; (3) searching and screening to identify essential studies; (4) describing and examining included studies; (5) describing, analyzing and synthesizing studies. Figure 2 shows the PRISMA step in reviewing articles about students' misconception in science.

In searching process, the researchers doing an investigation to some articles that are published in the scientific journal in the area of science education and indexed by the trustworthy institution to get data about the specific databases of students' misconception and diagnostic instruments. To analyze the matching studies in the articles, researchers conducted a specific search of some indexing institutions namely ERIC, EBSCO, SAGE, DOAJ, WILEY, JSTOR, ELSEVIER, SCOPUS, and WOS with document analysis approach. A restriction is used to focus some articles using English published from the year 2015 to 2019 to get the newest study of misconception articles in science education. Some stages of the process were held whereby every article was investigated and information of the studies was analyzed and discussed by two researchers. After reducing some articles from 1501 original articles investigated using the abstract and keyword search, the present study selected a total of 111 research articles which have a concern on misconception and diagnostics assessment.

Data from the 111 selected articles were analyzed using a form which recorded keyword information about the studies: (1) authors, (2) year of publication, (3) type of publication, (4) field study, (5) science concept, (6) view topic, (7) research instruments, (8) major findings. The first step of the reviewed studies was to analyze using descriptive statistics to find percentages of instrument used in current research. The next step for the questions concerning common misconception in science education, we analyzed the science concept or misconception topic of every article. For detail of the diagnostic instruments, we investigated and synthesized each material according to a group of the major diagnostics test used in studies categorized on the interview, multiple-choice tests, multi-tier tests, and open-ended tests.

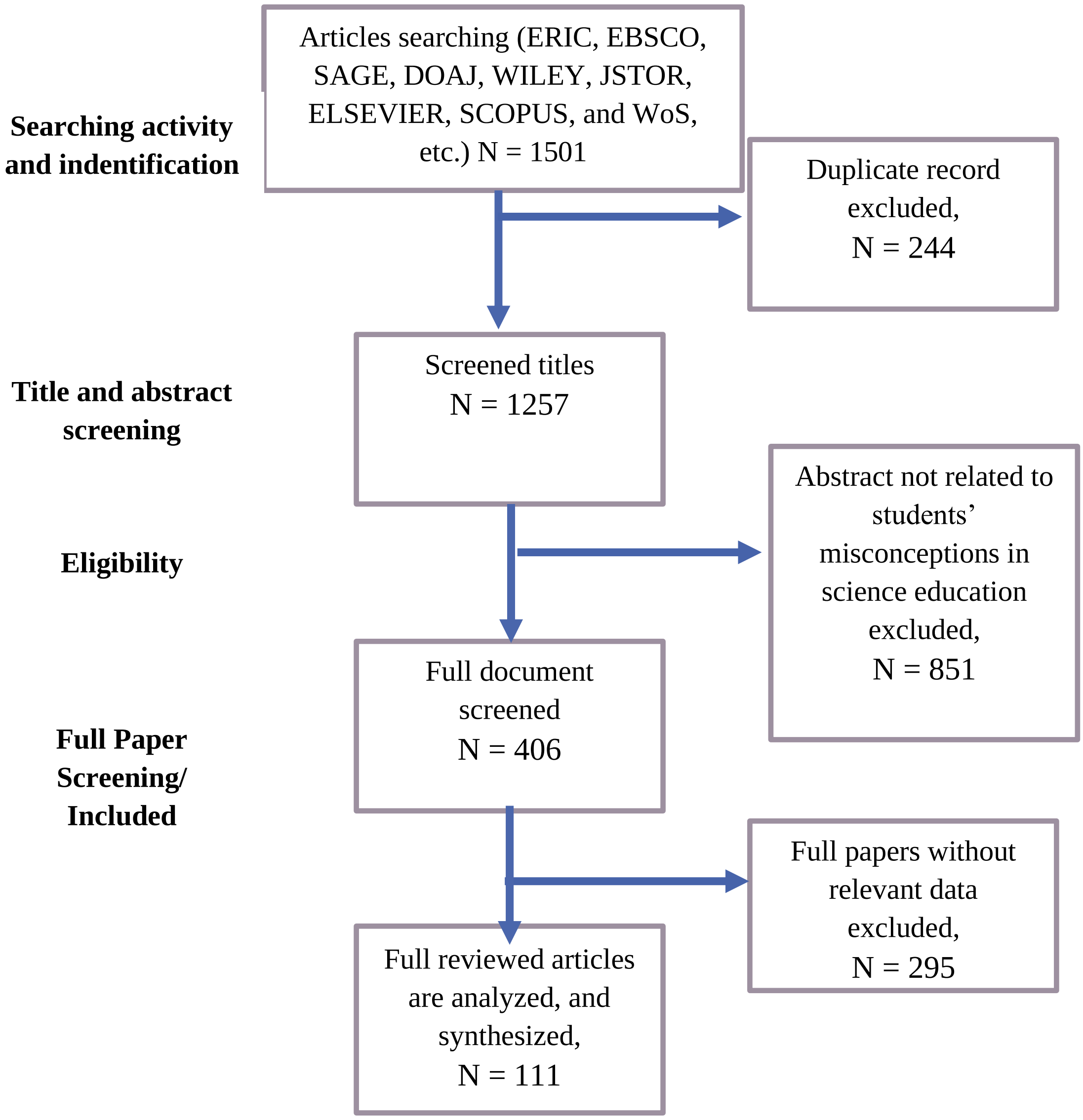

Searching activity and indentification
Articles searching (ERIC, EBSCO, SAGE, DOAJ, WILEY, JSTOR, ELSEVIER, SCOPUS, and WoS, etc.) N = 1501
Duplicate record excluded, N = 244
Title and abstract screening
Screened titles N = 1257
Eligibility
Abstract not related to students' misconceptions in science education excluded, N = 851
Full Paper Screening/ Included
Full document screened N = 406
Full papers without relevant data excluded, N = 295
Full reviewed articles are analyzed, and synthesized, N = 111

Figure 2. Flow diagram of information through different steps on the review process.

In the literature review process, the process is carried out repeatedly and gradually. Research articles were investigated based on abstracts, methods, instruments, and result of misconception analysis. The main discussion of diagnostic assessment in the articles is used as data instruments that are often used and compare strengths and weaknesses with each other. In conducting a literature review, researchers paid a specific interest to the type of multiple-choice instrument and multi-tier test because of the frequent use of this test. but this problem does not mean that other instruments like open-ended questions and interviews are not used in various studies, they are still used and have an influence on misconception in scientific analysis.

2.3.3 Instruments and related concepts in investigating students' science misconceptions.

To measure and identify students' misconception in several scientific concepts, various diagnostic tests have been developed and used. The interview, open-ended question, multiple-choice question, and multiple-tier test were found to be the most frequently used in science education research. However, each test has its advantages and disadvantages as discussed in several studies.

Table 1. Proportions of diagnostic instrument used to examine and identify science misconceptions.

Diagnosis method			Percentages (%)
Interviews			10.74
Open-ended questions Test			23.97
Multiple choices Test			32.23
Multiple-tier Test			33.06
	two-tier	9.92	
	three-tier	16.53	
	four-tier	4.13	
	multi-tier	2.48	
Total			100

Based on 111 studies included in this study, the most widely used diagnostic test was found as multiple-tier tests (33.06%). However, this study also found some scholarly works that use combination of some diagnostic assessment to get better result in research. We found that researchers usually add interviews as second instrument used to identify science misconception. Table 3 shows interview used as second instrument in some research, but the main instrument mostly using multi-tier test, open-ended questions test, and multiple-choice test. In Table 9. We found some studies that use some combination multi-tier test to identify misconceptions in science.

Table 1 shows the percentages of articles reviewed in this study, followed by other diagnostic tools such as multiple-choice tests (32.23%) and multiple-tier tests (33.06%), and open-ended tests (23.97%). However, every test has benefits and drawbacks over when used in assessing student conceptions, but studies are using multi-diagnostic tests (2.48%) which means that the study does not only use a single instrument but two or three types of diagnostic methods.

Table 2. Common misconception topics in reviewed articles.

Subject		
Physics	Chemistry	Biology
Photoelectric effect	Chemical bonding	Adaptations, habitat, biosphere, ecosystem, food chain and food web, functions of an ecosystem, biomass and biodiversity.”
Light	Electrolyte and Ion	Osmosis and diffusion
Impulse and momentums	Fire concept	Plant transport
Geometrical optics	Thermochemistry, chemical kinetic	Antibiotic resistance
Dynamics rotation	Carbohydrates	Acid rain, global warming, greenhouse effect, and ozone layer depletion
Simple current circuits	Enzyme interacts	Water cycle
Power	Electrochemistry	Photosynthesis
Radioactivity	The Mole Concept	Nature of science
Heat, temperature and internal energy	Acid-base	Digestive system
Static electricity	Ionic and covalent bonds concepts	Energy and climate change
Projectile motion	Acid-Base and Solubility Equilibrium.	Evolution of biology
Geometrical optics	Redox titration	Human reproduction
Fluid static		The human and plant transport systems.
Electrostatic charging		Global warming
Net force, acceleration, velocity, and inertia.		Ecological concepts
Lenses		
Heat, Temperature and Energy Concepts		
Newton’s law		
Temperature and heat		
Energy		
Sinking and floating		
Magnet		
Density		
Moon phase		
Gases		
Mechanics		
Astronomy		
Solid matter and pressure liquid substances		
Thermal physics		
Mechanics		
Hydrostatic Pressure and Archimedes Law		
Hydrostatic pressure concept		
Astronomy		

There are some factors that cause students' misconception in science namely, everyday life experiences, textbook, teacher, and language used, but we found that the reason for students has misconception in science because of characteristics for abstract and complex concept and difficulties to understand it, and usually find in everyday phenomenon. In example, for Light and Optics concepts, some studies show that Light and Optics are difficult to be understood by students. The characteristic of the complex abstract and abstract tends to lead student and teacher misconception and gives difficulties in conducting learning process. Light and Optics also easily found in everyday phenomenon which make student familiar with this topic and carrying their own understanding leading misconception in learning process (Ling, 2017; Widiyatmoko & Shimizu, 2018)

This study revealed that topics of physics, chemistry, and biology subject in science which commonly distributing misconception for the student are physics with 33 concepts, chemistry with 12 concepts, and biology with 15 concepts as shown in Table 2. Table 2 reveals that the most field topic in science caused misconception is physics subject.

Interview

Among several methods in diagnosing misconceptions, interviews have a very important role because researchers may get detailed information about students' cognitive knowledge structures. In fact, interviewing is one of the best and most widely used to find out the knowledge and possible misconceptions that students have (Fuchs & Czarnocha, 2016; Jankvist & Niss, 2018; Wandersee et al., 1994). Interviews can be used to translate student responses or answers to be analyzed and classified based on appropriate scientific conceptions (Shin et al., 2016). Several interview techniques have been used in previous studies such as interviews for remedial learning (Kusairi et al., 2017), Interviews can be used as individual and group (Fontana & Prokos, 2016), Interviews as a complement test of multiple-tier question (Linenberger & Bretz, 2015; Mutlu, & Sesen, 2015; Murti & Aminah, 2019). Aas et al. (2018) stated that the interview group has strength in developing ideas and processes of interaction with students.

The purpose of interviewing is not to get answers to questions, but to find out what students think, what is in the minds of students, and how students think about a concept (Seidman, 2006). Gurel et al. (2015) state that when the right interview is conducted, interviewing is the most effective way to reveal student misconceptions. They also suggest that using a combination of interviews and other tests like multiple-choice will make the research instrument better. Although the interview has many advantages in getting information, a significant amount of time is needed for and the researcher needs training to conduct interviews. Besides, interview bias may be found in research because data analysis will be a little difficult and complicated (Tongchai et al., 2009)

Table 3. Interview in science assessment.

Field	Misconception topics	References	Status
Physics	Radioactivity	(Yumuşak et al., 2015)	Major
	Fluid static	(Kusairi et al., 2017)	Complement
	Heat and temperature.	(Ratnasari & Suparmi, 2017)	Complement
	Light	(Wartono & Putirulan, 2018)	Complement
Chemistry	Electrolyte and ion	(Shin et al., 2016)	Complement
	Thermochemistry, chemical kinetic	(Mutlu, & Sesen, 2015).	Complement
	Enzyme Interacts	(Linenberger & Bretz, 2015)	Complement
	Chemical bonding	(Enawaty, & Sartika, 2015)	Complement
	Particulate nature of matter	(Kapici, & Akcay, 2016)	Complement
Biology	Acid rain, global warming, greenhouse effect, and ozone layer depletion	(Karpudewan et al., 2015)	Major
	Evolution of biology	(Putri et al., 2017).	Complement
	Natural science	(Murti, & Aminah, 2019).	Complement
	Global warming	(Fajarini et al., 2018)	Major

Table 6. depicts information about articles used the interview as an instrument to reveal students' misconception in science. As shown in Table 3. interviews are widely used as the second or complement test in research to reveal misconceptions, this may be due to researchers being unable to work with large samples when using interviews as the only test and avoiding bias in assessing and doing an interview.

Open-ended tests

in the interest of investigating students' conceptual understanding, the open-ended question is a diagnostic method that is often used to identify student understanding in science education. This method gives students the freedom to think and write their ideas, but it is a little complicated to evaluate the results or responses because the problems of using the language and students tend not to write their understanding in complete sentences (Baranowski, & Weir, 2015). Krosnick (2018) stated that the open-ended test has several advantages, namely helping students express their ideas, having an unlimited range for answers, minimizing in the answers given by students. However, it also has some drawbacks such as difficulty interpreting and analyzing student answers requires special skills for getting meaningful answers, some response answers may not be useful, bias answers may occur if students do not understand the topic of the question. Table 4 gives information about some reviewed articles from 2015 to 2019 using an open-ended test to investigate student misconception in science.

Table 4. Open-ended tests in science assessment.

Field	Misconception topics	References
Physics	Projectile motion	(Piten et al., 2017)
	Net force, acceleration, velocity, and inertia.	(Gale et al., 2016)
	Heat, temperature and energy concepts	(Celik, 2016; Ratnasari, & Suparmi, 2017)
	Lenses	(Tural, 2015)
	Newton’s Law	(Alias, & Ibrahim, 2016)
	Energy	(Lee, 2016)
	sinking and floating	Shen et al., 2017)
	Light and magnet	(Zhang & Misiak, 2015)
	electric circuits	(Mavhunga et al., 2016)
	Density	(Seah et al., 2015)
	General physics concept	(Armağan, 2017).
	Mechanics	(Foisy et al., 2015; Daud et al., 2015)
	Digital system	(Trotskovsky & Sabag, 2015)
	Newton's Third Law	(Zhou et al., 2016)
	Energy in five contexts: radiation, transportation, generating electricity, earthquakes, and the big bang theory.	(Lancor, 2015)
Chemistry	Particle position in physical changes	(Smith & Villarreal, 2015)
Biology	Particulate nature of matter	(Kapici & Akcay, 2016)
	Nature of science	(Leung et al., 2015; Wicaksono et al., 2018; Fouad et al., 2015)
	Digestive system	(Istikomayanti & Mitasari, 2017; Cardak, 2015)
	Energy and climate change	(Boylan, 2017)
	Biological evolution	(Yates & Marek, 2015)
	Biology concept	(Antink-Meyer et al., 2016)
	Introductory biology	(Halim et al., 2018)
	Ecological concepts	(Yücel & Özkan, 2015)

From Table 4, we can find out that the physics and biology subject in the article reviewed used open-ended questions frequently. The open-ended question applied to the concept of fundamental issues in science.

Simple multiple-choice test

To overcome difficulties in interview and open-ended question test in assessment, diagnostic multiple-choice tests can be used to assess student conception with large numbers of participants. This test is usually the main test given before conducting a random interview. The development of multiple-choice tests on students had made valuable contributions to research related to student misconceptions. The presence of multiple-choice tests can help researchers or teachers to find student misconceptions in their classrooms (Abdulghani et al., 2015). The results of student misconception studies are widely reported using multiple-choice tests. The validity evidence of this test is also strong (Haladyna & Downing, 2011). Based on the review results it is known that multiple-choice tests are chosen because they are valid and reliable, easy to do the scoring, easy to manage, can use conventional paper and pencil test making it easier for researchers to assess students' misconceptions of science. The researcher or teacher will get information about students' misconceptions and knowledge by using instrument diagnostics. When student misconceptions are identified, they can provide remedy related to improper conception with various teaching approaches. Some of the benefits of using multiple-choice tests over other instruments have been discussed by multiple authors (Azizoğlu, & Geban, 2016; Eshach et al., 2018; Milner-Bolotin, 2015; Önder, 2017). In summary, some of the benefits of multiple-choice tests are: (1) This test allows researchers to make coverage of various topics in a relatively short time. (2) Multiple-choice tests are versatile and can be used at different levels of instruction. (3) They are objective in assessing answers and being reliable. (4) they are easy and quick to do the scoring. (5) They are suitable for students who have a good understanding but poor to write. (6) They are suitable as items of analysis where various variables can be determined for the analysis process. (7) They are valuable in assessing student misconceptions and can be used on a large scale.

The main difficulty in multiple-choice tests is interpreting students' responses if items have not been carefully constructed (Antol et al., 2015). Researchers can develop test items with good deception based on student answer choices. Tarman & Kuran (2015) suggested combining interview and multiple-choice test as an ideal instrument to identify students' understanding in the assessment process.

Besides, multiple-choice tests also get criticism and have some weaknesses. Bassett (2016) and Chang et al. (2010) state that multiple-choice tests have various weaknesses as follows: (1) Guess answers can cause errors on variances and break down reliability est. (2) Answer choices do not provide insight and understanding to students regarding their ideas. (3) Students are forced to have one correct answer from various answers that can limit the ability to construct, organize and interpret their understanding. (4) Writing a good multiple-choice test is difficult.

Another criticism related to multiple-choice tests was revealed by Goncher et al. (2016). They stated that multiple-choice tests do not offer knowledge that includes students' ideas and also offer sometimes true answers for the wrong reasons. In other words, multiple-choice tests cannot distinguish true answers from the true reasons or true answers that have wrong reasons (Caleon & Subramaniam, 2010; Eryılmaz, 2010), so that errors may occur in the assessment of student misconceptions (Peşman & Eryılmaz, 2010). Vancel et al. (2016) conduct research in developing and analyzing the answers of multiple-choice tests. The results of the research indicate that the correct answers in the multiple-choice test do not guarantee the correct reason and assessment of the questions made. to cope with the limitations of multiple-choice tests. In various recent studies, a multiple-tiers test was developed.

Table 5. Simple multiple-choice conceptual tests in science assessment.

Field	Misconception topics	References
Physics	Light	(Milner-Bolotin, 2015)
	Energy and momentums	(Dalaklioğlu & Sekercioğlu, 2015)
	Fluid static	(Kusairi et al., 2017)
	Impulse and momentums	(Soeharto, 2016; Samsudin et al., 2015)
	Temperature and heat	(Madu & Orji, 2015; Asri et al., 2017)
	Sport physics	(Kartiko, 2018)
	Energy and force	(Nwafor et al., 2015)
	Newtons' Law	(Ergin, 2016)
	Electric circuits	(Sadler & Sonnert, 2016)
	Gases	(Azizoğlu, & Geban, 2016)
	Physical concept	(Wind & Gale,2015)
	Heat transfer	(Wibowo et al., 2016)
	Thermal physics	(Malik et al., 2019)
	Moon phase	(Saenpuk & Ruangsuwan, 2019)
	Energy material	(Wijayanti et al., 2018)
	Light	(Wartono & Putirulan, 2018)
	Heat concept	(Haryono, 2018)
	Solid matter and pressure liquid substances	(Handhika et al., 2018)
	Sound	(Eshach et al., 2018)
	Hydrostatic pressure and Archimedes law	(Berek et al., 2016)
Chemistry	Municipal chemistry	(Milenković et al., 2016)
	Chemical bonding	(Vrabec & Prokša, 2016; Enawaty & Sartika, 2015)
	Enzyme Interacts	(Linenberger & Bretz, 2015)
	Chemical bonding and spontaneity	(Ikenna, 2015)
	Electrochemistry	(Önder, 2017)
	Acid-base	(Sadhu et al., 2017; Sadhu, 2019))
	Acid-base and solubility equilibrium.	(Masykuri & Rahardjo, 2018)
Biology	Photosynthesis	(Orbanić et al., 2016)
	Evolution of biology	(Putri et al., 2017; Helmi et al., 2019)
	Natural science	(Subayani, 2016; Murti & Aminah, 2019)
	Global warming	(Fajarini et al., 2018)
	Ecology	(Butler et al., 2015)

There are several examples of using multiple-choice tests in research to identify student misconceptions in science education. Table 5 gives information about some articles using multiple-choice tests as a diagnostic instrument. Most physics subject studies are carried out using multiple-choice tests, from other tests it can be concluded that physics also ranks the top in the field of science where students often experience misconceptions.

Two-tier multiple-choice test

In general, the two-tier tests are diagnostic instruments with a first tier in the form of multiple-choice questions, and the second tier in the form of reasons that are compatible with multiple-choice sets on the first tier (Adadan & Savasci, 2012). Student answers are stated correctly when the answer choices of contents and reasons were given correctly. Distracters in two-tier tests are based on a collection of literature, student interviews, and textbooks. Two-tier tests are the development of a diagnostic instrument because students' reasons can be measured and linked to answers related to misconceptions. With two-tier tests, researchers can even find student answers that have not been thought of before (Tsui & Treagust, 2010). Adadan & Savasci (2012) also stated that two-tier tests make students easier to respond the question and more practical to be used by researchers in various ways such as reducing guesses, large-scale use, ease of scoring, giving explanations regarding student reasoning. Table 6 summarizes the two tests used for research about students' misconceptions in science.

Table 6. Two-tier multiple-choice tests in science assessment.

Field	Misconception topics	References
Physics	Power	(Lin, 2016)
	Radioactivity	(Yumuşak et al.,2015)
	Impulse and momentums	(Saifullah et al., 2017)
	Astronomy	(Kanli, 2015)
Chemistry	Fire Concept	(Potvin et al., 2015)
	Thermochemistry, Chemical Kinetics	(Mutlu, & Sesen, 2015).
	The Mole Concept	(Siswaningsih et al., 2017)
		(Widarti et al., 2017)
	Acid-base and argentometric titration	
	Redox titration	(Widarti et al., 2016)
Biology	Osmosis and diffusion	(AlHarbi et al., 2015)
	Plant transport	(Vitharana, 2015)
	Antibiotic resistance	(Stevens et al., 2017)

The study that provides a critique of the use of two-tier tests has been conducted by Gurel et al. (2017) in the discipline of physics, especially for geometrical optics. They say that two-tier tests may provide an invalid alternative concept, but it is uncertain whether student errors are caused by misunderstandings or words that are not needed in the test which causes the question to be too long to read. So another test in the form of a four-tier test needs to be developed. Another disadvantage related to two-tier tests revealed by Vitharana (2015) is that the choice of answers in two-tier tests can provide guidance to students regarding the correct answers. Because the answer choices related to misconceptions has a logical relationship with the reason, for example, students can choose answers to the second tier because the answers must be related with responses to first-tier questions, or part of the two-tier test can provide answers that are interrelated and half correct, so students find it easier to find the right answer using this logic (Caleon & Subramaniam, 2010a). Therefore, two-tier tests may overestimate or underestimate student conceptions so that it is difficult to predict disparities in terms of student misconceptions and knowledge with two-tier tests (Caleon & Subramaniam, 2010a, 2010b; Peşman & Eryılmaz, 2010). To overcome this problem, an alternative blank answer is given in the part of the reason in the second-tier question so students can write responses that give explanations related to their understanding (Eryılmaz, 2010; Kanli, 2015; Peşman & Eryılmaz, 2010).

To sum up, two-tier tests have benefits compared with simple multiple-choice tests, interviews, and open-ended tests. This test provides an answer option for multiplying student reasoning or interpretation toward the question of misconception in science. However, two-tier tests have several limitations and disadvantages in distinguishing misconceptions, mistakes or scientific understanding. For this reason, several recent studies have conducted a three-tier and four-tier test to diagnose student misconceptions in science learning.

Three-tier multiple-choice test

The limitations that appear in two-tier tests encourage researchers to develop third tier tests that have items to measure the level of confidence in the answers given to each two-tier item question (Aydeniz et al., 2017; Caleon & Subramaniam, 2010; Eryılmaz, 2010; Sen & Yilmaz, 2017; Sugiarti, 2015; Taslidere, 2016). In the first three-tier tests, tests in the form of simple multiple-choice, at the second level in the form of multiple-choice with a choice of reasons and in the third level questions made using the level of confidence scale on the two previous levels of questions. Students' answers to each question item are considered correct when correct on the questions related to the concept and the reasons given with advanced confidence. Likewise, for students' answers which are considered wrong when the answer to the wrong concept choice is accompanied by wrong reasons that have a high level of confidence. Three tier tests are considered more accurate in identifying students' misconceptions. Because the Three-tier test can detect students' lack of understanding by using a level of confidence in the answers given by students, this condition helps researchers get a more accurate percentage of misconceptions that are free of doubt and lack of understanding of the concept because each student needs different treatments to correct their misconceptions.

Table 7. Three-tier multiple-choice tests in science assessment.

Field	Misconception topics	References
Physics	Photoelectric effect	(Taslidere, 2016)
	Heat and Temperature	(Kusairi, & Zulaikah, 2017; Putri & Rohmawati, 2018)
	Dynamics Rotation	(Syahrul, 2015)
	Simple Current Circuits	(Osman, 2017)
	Heat, temperature and internal energy	(Gurcay & Gulbas, 2015)
	Geometrical Optics	(Taslidere & Eryilmaz, 2015)
	Particulate Nature of Matter	(Aydeniz et al., 2017)
	Heat	(Irsyad et al., 2018)
	Kinetic theory of gases	(Prastiwi et al., 2018)
	Newton's Laws of Motion Concept	(Sulistri & Lisdawati, 2017)
	Hydrostatic pressure concept	(Wijaya et al., 2016)
	Astronomy	(Korur, 2015)
Chemistry	Chemical Bonding	(Sen & Yilmaz, 2017; Sugiarti, 2015)
	Carbohydrates	(Milenkovic, et al., 2016)
	Ionic and Covalent Bonds Concepts	(Prodjosantoso & Hertina, 2019)
Biology	Adaptations , habitat, biosphere, ecosystem, food chain and food web, functions of ecosystem, biomass and biodiversity”	(Oberoi, 2017)
	Human Reproduction	(Taufiq, et al., 2017)
	The Human and Plant Transport Systems.	(Ainiyah, et al., 2018)

In many uses of the three-tier test, researchers developed it by combining various diagnostic methods for misconceptions such as open-ended tests and interviews. The diversity of ways in collecting data related to student misconceptions provides a good foundation in the development of valid and reliable diagnostic assessments. Table 7 provides information on the use of three-tier tests to find out student misconceptions in science education. To sum up, three-tier tests have several advantages, which can determine students 'misconceptions more accurately because they can distinguish students' misconceptions and ignorance. Therefore, three-tier tests are considered more valid and reliable in assessing student misconceptions than simple multiple-choice and two-tier tests (Aydeniz et al., 2017; Taslidere, 2016). However, three-tier tests also have drawbacks because the level of confidence is only used in choices related to reasons so that there may be the overestimation of the proportions of knowledge in the student's answer scoring. For this reason, four-tier tests that provide a level of confidence in the content and reason are made and introduced recently.

Four-tier multiple-choice test and multi-tier test

Although the three-tier tests are considered to measure students' misconceptions free from errors and lack of student knowledge in a valid and reliable path, the three-tier tests still have some disadvantages due to limitations in converting confidence ratings on the first and second tier questions in the test. This situation causes two problems. First, the percentage of knowledge is too low, and both estimates are too excessive on scores of student misconceptions and correct answers.

Table 8. Four-tier multiple-choice tests in science assessment.

Field	Misconception topics	References
Physics	Geometrical optics	(Gurel et al., 2017; Fariyani et al., 2017)
	Energy and momentum	(Afif et al., 2017)
	Static electricity	(Hermita et al., 2017)
	Solid matter and pressure liquid substances	(Ammase et al., 2019)
Chemistry		
Biology		

In several self-reviewed articles related to students' misconceptions in science education, only a few studies are using four-tier tests rather than three-tier tests. Table 8 shows that the use of four-tier multiple-choice tests is only used in research in the field of physics. Although four tier multiple-choice tests are considered to be able to eliminate the problems mentioned in the previous tests, this test still has some drawbacks. There are requiring a long time for testing process, difficult to use in achievement tests (Caleon & Subramaniam, 2010), and the possible choice of students' answers at the first level can influence responses at the next tier questions (Sreenivasulu & Subramaniam, 2013; Ammase et al., 2019).

Table 9. Multi-tier multiple-choice tests in science assessment.

Field	Misconception topics	References
Biology	Concept of adaptation	(Maier et al., 2016)
	Water Cycle	(Romine et al., 2015)
	Concept of water characteristics.	(Sari, 2019)
Chemistry		
Physics		

We also found three studies that tried to combine several multi-tier questions into new multiple-tier questions (Maier et al., 2016; Romine et al., 2015; Sari, 2019). The instrument test used is a combination of two-tier, three-tier and four-tier question. Table 9 shows that the use of multiple tier tests is still rarely done in science education.

In the last part of discussion, this study will give some comparisons related to the trends of diagnostic instruments used to identify students’ misconception in science from Wandersee et al., (1994), Gurel et al., (2015), and this study highlights the benefits and drawbacks of each test instrument used in diagnostic research on science education.

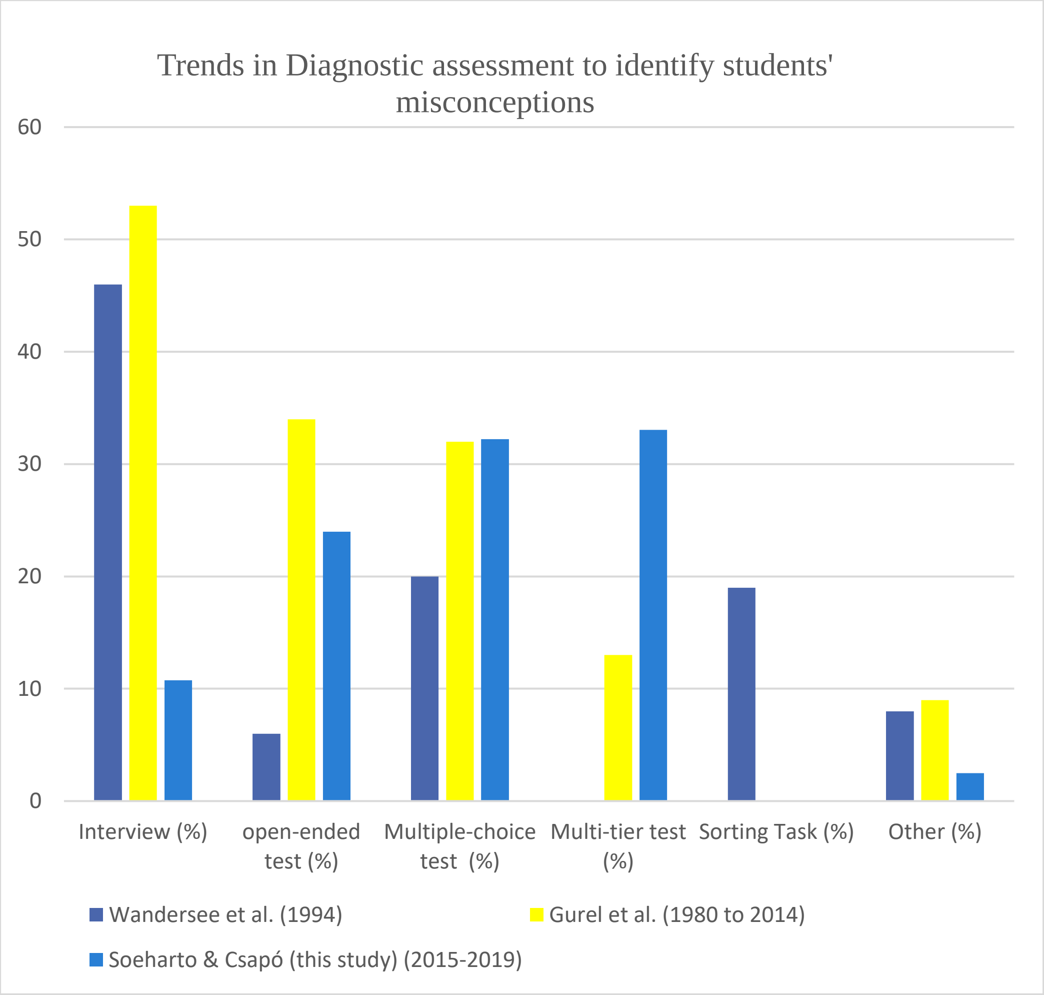


Figure 3. Trends in diagnostic assessment to identify students' misconception in science**.**

2.4 Student misconceptions and the importance of research for science education

Student misconceptions had been a problem in the science education area for 30 years ago. Driver and Easley (1978) had pointed out there are a conceptual understanding among adolescent related to science concepts well known as ”student misconceptions”. Many studies were related to student misconceptions in learning science because the characteristic misconceptions in science are resistant to change, persistent, and rooted in some science concepts (Boone et al., 2013; Greiff et al., 2018; Morrison et al., 2019; Topalsan & Bayram, 2019). Besides, if students experience misconceptions in learning science, students will find it difficult to learn science at a higher level. Student misconceptions in science can lead students to get low academic performance scores for science education subjects such as physics, biology, and chemistry.

Although many studies are related to student misconceptions in science concepts across disciplines, only a few studies focus on understanding the inherent difficulty level of items in science concepts in various science disciplines (e.g., Liu et al. (2015); Park and Liu (2019)). Recently, Lancor (2015) and Chen et al. (2014) found that students' understanding of science concepts is different for each discipline, which implied the importance of understanding the difficulty level of items in science concepts across science disciplines. Students must be able to develop their understanding of scientific concepts across all disciplines to achieve the success of the learning objectives (Krajcik et al., 2014). This finding proves that the level of difficulty in scientific concepts will hinder the development of students' understanding in learning. Knowing science concepts embedded in various disciplines is necessary to investigate students' strengths and weaknesses against different scientific concepts so that teachers can have the empirical evidence required to teach science concepts across the science disciplines better.

2.5 Pre-service science teacher misconceptions

Studies involving preservice science teachers (PST) or university students have shown that misconceptions in science occur throughout the different education levels, even among senior teachers or professional teachers (Becker & Cooper, 2014; Duit, 2014; Kiray & Simsek, 2021; Laliyo et al., 2019; Liampa et al., 2019; Stefanidou et al., 2019). Kaltakci-Gurel et al. (2017) found that PST sometimes share misconceptions that students hold in their knowledge. These misconceptions exist in learning design and learning activities, which directly reinforce students' misconceptions instead of correcting them. In Indonesia, science teachers have a special agenda called 'remediation' to correct students' misconceptions. Remediation activities are usually held after students' examinations in science disciplines, where science teachers reconstruct students' knowledge regarding science concepts. Galvin and Mooney (2015) highlighted the importance of identifying misconceptions of PST, undergraduate students in teacher training and education majors to improve the quality of science teachers and reduce student misconceptions. If science teacher misconceptions are not corrected, science teachers may fail to properly teach science concepts to students in their learning activities (Arslan et al., 2012; Gurbuz, 2015).

2.6 Instruments for identifying misconceptions in science

Student misconceptions are difficult to identify with traditional methods. Educators have to revise and identify student misconceptions to help students understand new concepts and finally provide opportunities for students to apply these concepts to science problems (Butler et al., 2015). To evaluate and identify students' basic knowledge of concepts in science, researchers used a diagnostic test. The diagnostic test assesses students' proportional knowledge on the basis of the science content, the science teacher can develop a clear idea about the nature of the students' knowledge by using a diagnostic test at the beginning or the end of the learning activity (Peterson et al., 1989; Taslidere, 2016; Treagust, 1986).

Researchers in science majors have used and developed numerous instruments to assess student misconceptions or student conceptual understanding (Soeharto, Csapó, et al., 2019). Two-tier multiple-choice diagnostic tests are the most reliable assessment tool developed to identify student misconceptions in science education majors because the multiple-choice test merely assessed student content knowledge without considering the reasoning behind students' responses (Chabalengula et al., 2012; Gurel et al., 2015). In a two-tier multiple-choice test, the first tier assesses students' insight about science concepts, whereas the second tier investigates student reasoning for their choices in the first tier. However, the two-tier multiple-choice test cannot differentiate students' mistakes due to lack of knowledge or simply guessing answers (Caleon & Subramaniam, 2010; Chabalengula et al., 2012). Thus, scholars introduced having the Certainty Response Index (CRI) embedded in the question, which measures the respondent level certainty in the first two tiers, and they call this test the three-tier multiple-choice diagnostic test (Gurcay & Gulbas, 2015; Peşman & Eryılmaz, 2010). However, regardless of the students having right or wrong answers, the answers with a low level of confidence were categorized as a lack of knowledge, and wrong answers with a high level of confidence were categorized as a misconception (Kaltakci-Gurel et al., 2017; Peşman & Eryılmaz, 2010). Instead, of using the confidence level choices or CRI on a three-tier or four-tier multiple-choice diagnostic test to differentiate between students' guessed answers or lack of knowledge answers, this study tries a new approach to analyze items: two-tier multiple diagnostic tests using an objective instrument based on Rasch measurement. The Rasch measurement was chosen because this analysis can provide accurate results of the level of student ability and the difficulty of items, even analyzing the likelihood of students just guessing the answers

(Sumintono & Widhiarso, 2014).

2.7 The development of the two-tier multiple choice test to assess misconception in science

In recent years from 2015 to 2019, multi-tier diagnostic tests are a popular assessment tool developed to identify student misconceptions in various research areas (Soeharto, Csapó, et al., 2019). The two-tier test is the first example in the development of a multi-tier test to diagnose student misconceptions. The two-tier multiple-choice test consists of first-tier and second-tier. The first tier assesses student conceptions, and the second tier assesses student reasonings without confident levels (Adadan et al., 2012; Korkmaz et al., 2018). We constructed item in first tier based on student common misconceptions in science. The first-tier question will evaluate student content knowledge. The second tier was constructed based on possible student reasoning related to scientific conception and possible alterative conceptions. The student answer is scored if the student can answer the content and reason correctly. Two-tier tests were developed as a diagnostic instrument because student conceptions and reasons are linked to understanding scientific misconceptions. Researchers can even find student answers with two-tier tests that have not been thought of before with blank option choice (Tsui & Treagust, 2010). Students are more accessible in responding to the question, and this test is used practically by researchers in various ways, including large-scale use, ease of scoring, and explanations regarding student reasoning (Adadan et al., 2012).

On the other hand, there are criticisms regarding the use of two-tier tests in identifying misconceptions. Gurel et al. (2015), in his research that identified misconceptions of geometrical optics in physics subject stated that two-tier tests might produce invalid misconceptions due to a lack of level of uncertainty where the researcher cannot ensure that the student's answer is the correct answer to guess, misconception, or concept. Although there are weaknesses in measuring student misconceptions because they cannot confirm students' answers with the confidence tier as in the three-tier and four-tier tests, the weaknesses in the form of guess answers, confident level issues, and missing data on the two-tier can be overcome by running the Rasch measurement model.

2.8 Inductive reasoning

Inductive reasoning may be defined as the cognitive activity of generating inferences that meet two criteria: direction and confidence level. In relation to direction, students move from specific observation cases to formulate general principles. With regard to confidence level, students start reasoning from a position of uncertainty to form related hypotheses (Feeney & Heit, 2007; Perret, 2015). Inductive reasoning is a form of reasoning, which can be broadly defined as the process of drawing conclusions that aim to solve problems and arrive at decisions (Lee et al., 2021; Sternberg et al., 2012). Inductive reasoning is concerned with deriving logically sound conclusions from a collection of premises (Feeney & Heit, 2007). In reasoning, one starts from the known to reach and/or evaluate a new conclusion (Sternberg et al., 2012). Inductive reasoning is the process of applying prior information to generate predictions about new cases (Hayes & Heit, 2017). Numerous interpretations of inductive reasoning can be found in various disciplines, including mathematics, philosophy, and psychology. Inductive reasoning is generally considered to be a cognitive process to enable one to generalize the rules from initial observations to arrive at a general conclusion (Csapó, 1997; Stephens et al., 2020).

Inductive reasoning plays a vital role in various cognitive activities such as feature attribution, analogical reasoning, causal reasoning, and probabilistic judgment. Furthermore, it is considered a pivotal element for understanding knowledge on a regular basis and for determining concepts and categories in daily activities (Klauer, 1996). In the inductive process, hypothetical rules to solve unfamiliar problems that can be tested on further action and observation are generated (Perret, 2015). In essence, inductive reasoning plays a role in understanding various knowledge and the application thereof to solve unfamiliar cases. Furthermore, it is included as one of seven factors in mental abilities that describe individual intelligence (Csapó, 1997; Kinshuk et al., 2006; Nikolov & Csapó, 2018; Perret, 2015)

2.9 The role of inductive reasoning in student development

Inductive reasoning can predict fluid intelligence and crystallized intelligence, which refers to students' ability to solve new problems in working memory (Feeney & Heit, 2007; Perret, 2015). Strobel et al. (2019) revealed that fluid intelligence can be measured by utilizing an inductive reasoning test. Additionally, student inductive reasoning was also defined as students' ability to elaborate on various insights in one's long-term memory (Perret, 2015). In turn, students' inductive reasoning can affect intelligence in similar ways. Inductive reasoning can be employed to solve new problems and support strategies in solving the same problems in different contexts (Feeney & Heit, 2007). Several studies have also demonstrated that inductive reasoning is an essential predictor of academic achievement and science performance (Van Vo & Csapó, 2020, 2021). In this study, inductive reasoning was assumed as a construct or latent factor, and it can be tested empirically using the inductive reasoning test. In the learning context, inductive reasoning plays a vital role in facilitating the learning process because inductive reasoning ability can help a student solve a complex problem. Therefore, assessing students' inductive reasoning is more profitable than assessing intelligence in the field of education. Many studies have also provided evidence that the higher the students' inductive reasoning ability, the higher their ability in various fields of science such as natural science, mathematics, attitudes, and languages (Childers & Exemplars, 2020; Kambeyo & Wu, 2018; Lee et al., 2021; Nikolov & Csapó, 2018; Sosa-Moguel & Aparicio-Landa, 2021; Van Vo & Csapó, 2021).

The primary objective of several inductive reasoning studies has focused on gender and grade at school. The majority of studies on student inductive reasoning have examined gender differences. Therefore, findings related to the effect of gender differences on student inductive reasoning are inconsistent in relation to the particular context and culture. Some studies have revealed that the inductive reasoning ability of male students is superior to that of female students (Strobel et al., 2019; Venville & Oliver, 2015). On the contrary, Díaz-Morales and Escribano (2013) found that female students' inductive reasoning abilities were superior to that of male students in predicting school achievement. Several studies also concluded that there were no significant gender differences between females and males in assessing inductive reasoning (Molnár & Csapó, 2011; Kambeyo & Wu, 2018; Kinshuk et al., 2006; Sosa-Moguel & Aparicio-Landa, 2021). Grade levels or age groups also affected students' inductive reasoning. As noted previously, Van Vo and Csapó (2020) demonstrated that students' inductive reasoning ability tended to increase regularly among 5th, 7th, 9th, and 11th grade students in Vietnam. While students' inductive reasoning tended to improve gradually from the 3rd grade (8–9 year-olds) to the 11th grade (16–17 year-olds), those in the 7th grade exhibited rapid development (Csapó, 1997; Díaz-Morales & Escribano, 2013; Molnár & Csapó 2011; Pásztor et al., 2018; Sosa-Moguel & Aparicio-Landa, 2021). No studies have been conducted on the effect grade has on inductive reasoning in Indonesia.

2.10 Assessing inductive reasoning in the educational context

In a meta-analysis, Waschl and Burns (2020) revealed that 40 different test types had been employed to measure inductive reasoning. The most common inductive reasoning test was designed to measure reasoning problems related to series completion, analogies, geometric matrices, and classification (Csapó, 1997; Hayes & Heit, 2017; Nikolov & Csapó, 2018; Van Vo & Csapó, 2020; Waschl & Burns, 2020; Wu & Molnár, 2018). Series completion tasks require students to determine the relationships in a given completion series such as numbers, letters, objects, and words. Students can solve a series completion task if they can find the relation between the given components so as to determine the next component as a solution (Klauer & Phye, 2008; Leighton & Sternberg, 2003; Waschl & Burns, 2020). The analogy task, which involves the structure of a display on an object such as figures, numbers, and letters, can be solved by assessing the sample information in the task. This task is frequently used to measure students' intelligence (Hotulainen et al., 2018; Klauer & Phye, 2008; Stephens et al., 2020; Strobel et al., 2019; Venville & Oliver, 2015). The classification task involves combining various forms of problems that comprise words, figures, and numbers that require students to identify answers that are unrelated to the others. The geometric matric task, in which a set of images is provided in a matrix where the rows and columns have particular rules, requires students to determine relationships and information to find missing images (Csapó, 1997; Klauer, 1996; Klauer & Phye, 2008; Waschl & Burns, 2020). Klauer and Phye (2008) formulated the genealogy of tasks in inductive reasoning to create inductive reasoning items so as to help researchers assess students' inductive reasoning. In Figure 4, Klauer's diagram that depicts the genealogy of tasks in inductive reasoning is portrayed. In this dissertation, the instrument was adapted and translated from the inductive reasoning test developed by Csapó (1997) and (Pásztor, 2016) for Indonesian purposes.

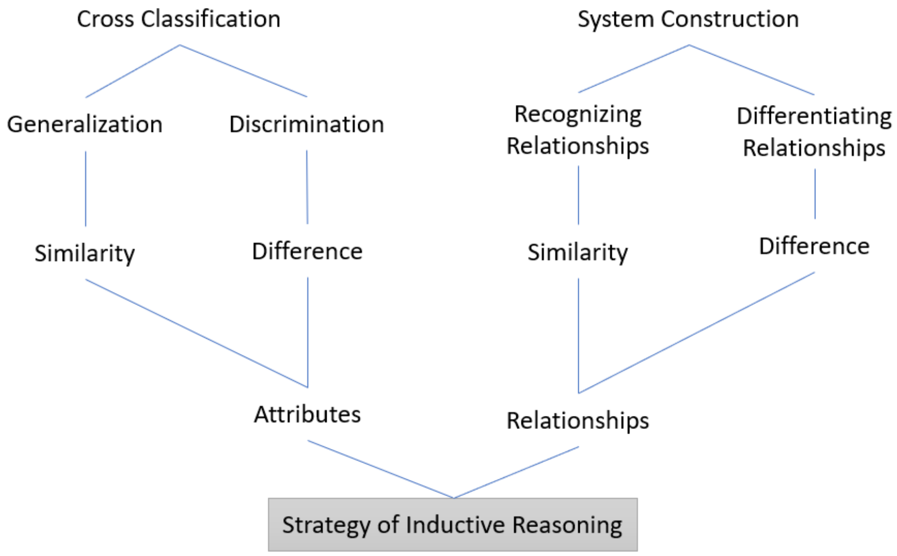


Figure 4. The genealogy of tasks in inductive reasoning (Klauer & Phye, 2008).

2.11 Possible relationship between students' misconceptions, science achievement, and inductive reasoning.

Based on preliminary literature review, there are no studies that specifically investigate the relationship between science misconceptions and inductive reasoning skills. However, the connection between them can be predicted indirectly with student achievement or performance in science subject. Several studies have investigated the relationship between science achievement and science misconceptions. Fuentes (2021) found a significant difference in science performance when grouped by the level of science misconceptions. Baweja (2017) reported a significant relationship between scores on student misconceptions and achievement in science. Students was also indicated having various misconceptions and other difficulties in science concepts that affecting their science performance (Eshach et al., 2018; Samsudin et al., 2021). A major concern should be given to differences between the conceptions of science and the conceptions of students from advanced science courses to improve science achievement (Abimbola, 1988; Subali et al., 2019).

In other hands, several researches have shown that inductive reasoning can help students overcome student achievement, which are a major challenge in science education (Bao et al., 2009; Korom et al., 2017; Molnár et al., 2013; Van Vo & Csapó, 2023, 2023). Inductive reasoning also had a significant effect on STEM achievement across grade levels (Van Vo & Csapó, 2023). Reference [10] reported a significant positive relationship between attitude towards science and reasoning ability in science of higher secondary students. Lin and McNab (2006) stated inductive reasoning ability as the best predictor for academic performance in science. Maryanto (2019) found that there is a difference in science instruction through deductive and inductive reasoning on improving learning achievement. Therefore, we can assume that inductive reasoning can be an important factor to predict student achievement in science.

In order to avoid misconceptions in the classroom, analogies and the process of analogic reasoning are essential. It has been demonstrated that the use of analogies improves conceptual teaching and learning, as well as the detection and eradication of misunderstandings (Harman & Çökelez, 2017). However, it is important to remember that analogies have two sides and might lead to misunderstandings if they are not fully understood (Sholihah et al., 2021). Therefore, educators must carefully choose analogies that are well-known to their students, directly link the parallel to the desired topic, and clear up any misunderstandings by pointing out the analogy's shortcomings (Ancker & Begg, 2017). Analogies can also be used as a diagnostic tool for evaluation, revealing misunderstandings pupils may have as well as the foundational knowledge underlying such misconceptions (Fotou & Abrahams, 2020). The number of misconceptions students have about various ideas has been shown to be reduced by the use of the bridge analogies teaching technique (Yilmaz et al., 2006). But it is crucial to thoroughly look at the elements that lead to the formation of analogical misconceptions (Zook & Maier, 1994). Misconceptions about particular science concepts such as matter conservation have been addressed using analogous instruction (Zook & Maier, 1994). Additionally, using an analogy-based approach in educational comics in Indonesian context can aid in the development of students' scientific notions by stimulating deeper understanding and guiding analogical thinking (Hesti, 2021). Overall, analogies and analogical reasoning can be valuable aspects for preventing misconceptions in the classroom, but their application must be done so carefully and effectively.

In conclusion, misconceptions in science education can be raised from inductive reasoning and science achievement. Even though the empirical research in this dissertation not covering the modelling research between inductive reasoning and misconceptions in science, the future research can be implemented because both factors have been assessed comprehensively in this dissertation.

CHAPTER 3. METHODOLOGY OF EMPIRICAL STUDIES

3.1 Research aims and research questions

3.1.1 Research questions for study 1

Study 1 investigates and evaluates the psychometric properties of the developed instrument, examines student misconceptions in science learning, and identifies background factors affecting student misconceptions in the learning context. The following are the research questions below.

RQ1.1: Does the developed instrument achieve reliability and validity based on Rasch measurement?

H1.1: We hypothesized that the two-tier multiple choice diagnostic test achieved the acceptable threshold based on Rasch measurement.

RQ1.2: How do items and persons interact in the developed instrument?

H1.2: It is expected that there is interaction between items and person based on Wright map (Planinic et al., 2019; Sukarelawan et al., 2021).

RQ1.3: How do the student misconceptions develop in science learning?

H1.3: We hypothesized that student misconceptions did not have significant differences across grade level (Kaltakci-Gurel et al., 2017; Mintzes et al., 2005; Tsui & Treagust, 2010; Wandersee et al., 1994).

RQ1.4: Is there an instrument bias based on gender according to differential item functioning (DIF)?

H1.4: No DIF issue is expected based on gender for the two-tier multiple choice diagnostic test (Wyse & Mapuranga, 2009).

RQ1.5: How do students' misconceptions develop across school grades?

H1.5: We hypothesized that students' misconceptions in upper grade higher than lower grade (Butler et al., 2015; Laliyo et al., 2019).

RQ1.6: What are the factors predicting student misconceptions in science?

H1.6: We assumed that gender and grade level are latent predictors for student misconceptions in science.

3.1.2 Research questions for study 2

Study 2 investigates item difficulty patterns, item–person map interaction, and DIF based on gender and grade across science disciplines using the two-tier multiple-choice diagnostic test for assessing student misconceptions. Hence, we set out the following research questions.

RQ2.1: Are the items on the instrument used valid and reliable?

H2.1: We expected the instrument used in main study hold acceptable validity and reliability same as in pilot study (Soeharto, 2021).

RQ2.2: What are the item difficulty patterns measured by diagnostic instruments for assessing student misconceptions on science concepts?

H2.2: We hypothesized there are differences in item difficulty patterns based on science concepts (Park & Liu, 2019).

RQ2.3: To what extent are the item difficulties able to describe the concepts that cause students misconceptions across disciplines and science concepts?

H2.3: We hypothesized there are differences in item difficulty patterns based on disciplines, science concepts, and interaction between science concepts and disciplines (Park & Liu, 2019).

RQ2.4: Are there any DIF issues based on gender and grade?

H2.4: No DIF issue is expected based on gender and grade (Sukarelawan et al., 2021; Wyse & Mapuranga, 2009)

3.1.3 Research questions for study 3

Study 3 explores student misconceptions in science concepts across school grades, examine student–item interaction regarding science concepts, detect outliers in student misconceptions and predict background factors that influence students' misconception in sciences. The following are the research questions below.

RQ3.1: Did the students provide guesses or inconsistent answers (i.e. misfitting persons) as their science misconceptions were assessed?

H3.1: We hypothesized that there are guessing or inconsistent answers in dataset.

RQ3.2: Are the items on the instrument used valid and reliable?

H3.2: We expected the instrument in main study after excluding misfitting persons hold acceptable validity and reliability same as in pilot study (Soeharto, 2021).

RQ3.3: How did students and items interact based on the person–item map and grade level?

H3.3: We hypothesized that students and items have interactions based on grade level.

RQ3.3: To what extent does the collected data fit the Rasch model and Confirmatory factor analysis models?

H3.3: We hypothesized that fit criteria achieved according to Rasch model and confirmatory factor analysis.

RQ3.4: How do students' science misconceptions differ in terms of gender and grade level?

H3.4: We expected that no differences in term of gender and grade level (Kaltakci-Gurel et al., 2017; Mintzes et al., 2005; Tsui & Treagust, 2010; Wandersee et al., 1994).

RQ3.5 Which factors predict student conceptions in science?

H3.5: We assume that grade level can predict student conceptions in science (Butler et al., 2015; Gurbuz, 2015; Soeharto, 2021).

3.1.4 Research questions for study 4

Study 4 assesses the adapted Indonesian version of the inductive reasoning test by determining its validity and reliability so as to evaluate Indonesian students' inductive reasoning and to classify their inductive reasoning levels in accordance with grade and gender. The following are the research questions for this study.

R4.1: In accordance with Rasch parameters, what is the reliability and validity of the adapted inductive reasoning test?

H4.1: We hypothesized that psychometric properties of Indonesian the adapted inductive reasoning test are acceptable (Korom et al., 2017; Molnár et al., 2013; Pásztor, 2016; Van Vo & Csapó, 2023)

R4.2: Is DIF detected between paper-based and online-based tests?

H4.2: No DIF issue is expected based on test method (Csapó et al., 2019; Csapó & Molnár, 2019)

R4.3: Is there any DIF based on gender and grade levels? No DIF issue is expected based on gender and grade (Sukarelawan et al., 2021; Wyse & Mapuranga, 2009)

H4.3: No DIF issue is expected based on gender and grade in main study after excluding outlier (Sukarelawan et al., 2021; Wyse & Mapuranga, 2009)

R4.4: How is the evaluation of Indonesian students' inductive reasoning across grade and gender?

H4.4: We assume that there are differences in result from the evaluation of Indonesian students' inductive reasoning across grade and gender(Van Vo & Csapó, 2020).

R4.5: What is the classification of the difficulty of inductive reasoning items and students' inductive reasoning abilities when employing Rasch analysis?

H4.5: We hypothesized that students' inductive reasoning abilities classified into moderate categories.

3.2 Instruments

3.2.1 Background questionnaires

The background questionnaire was adapted from the Indonesian version's PISA 2015 SES questions (OECD, 2016). The questionnaire is embedded in the developed multi-tier diagnostic test body in the online and paper-based format. The background questionnaire in this study consists of information such as gender, parents' level of education, parents' jobs, and student performance in the science subjects in the previous semester. The background questionnaires were functioned to depict demographic profile and to evaluate predictors that affect student misconceptions in science using stepwise regression analysis.

3.2.2 The two-tier multiple choice diagnostic test

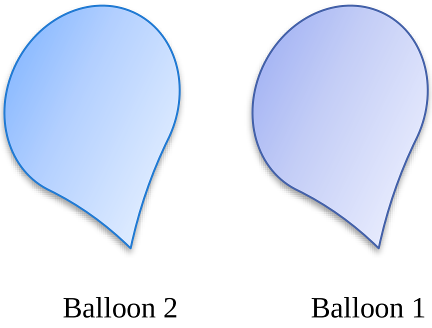
To identify students' misconceptions in science, 32 item questions were developed and divided from three science subjects, physics, biology, and chemistry (See Appendix 4). Sixteen concepts distributing misconceptions in science selected were shown in Table 2. Concepts and item numbers in the developed two-tier multiple-choice diagnostic test. In identifying common misconceptions in science, we investigated literature review studies and misconceptions in science handbooks (AAAS, 2019; Allen, 2014; Csapó 1998; Soeharto, et al., 2019). Then the selected concepts had been adjusted according to Indonesian education curriculum the Curriculum 2013, especially on the senior high school level. The developed test consists of two tiers. The first tier will represent student conceptions in science that are linked to the question, and the second tier will represent student justifications for those conceptions. In the event that the reasoning choice is not available, we also offer a blank option to give students a chance. For each of the items, a right response received 1 point, while an incorrect response received 0. If a student completes the first and second tiers of the task properly, they receive 1 point. All items in the test were translated using the back-forward translation from English to Indonesian and then from Indonesian to English by researchers. Table 10 present science concepts and indicators in the developed two-tier multiple choice diagnostic test. Table 11 depicts a sample item in Physics used in Indonesian an English version.

Table 10. Science concepts and topics in the developed two-tier multiple-choice diagnostic test (N=32).

No	**Topics and Concepts**	**Item Number**
1	Topic: Kinetic energy Concept: Kinetic energy is associated with the speed and the mass of an object	1, 2
2	Topic: Thermodynamics – Thermal energy Concept: Thermal energy is associated with the temperature and the mass of an object and the material of which the object is made.	3
3	Topic: Thermodynamics – Thermal energy Concept: Thermal energy of an object is associated with the disordered motions of its atoms or molecules and the number and types of atoms or molecules of which the object is made.	4
4	Topic: Impulse and Momentums Concept: Impulse and momentums relation	5,6
5	Topic: Atoms and molecules Concept: When heated, solids can change into liquids and liquids can change into gases. When cooled, gases can change into liquids and liquids can change into solids. These changes of state can be explained in terms of changes in the proximity, motion, and interaction of atoms and molecules.	7
6	Topic: Atoms and molecules Concept: For any single state of matter, increasing the temperature typically increases the distance between atoms or molecules. Therefore, most substances expand when heated.	8
7	Topic: Forces Concept: Heavy objects do not fall at a greater speed than light objects.	9
8	Topic: Forces Concept: An object at rest condition has forces acting upon it	10
9	Topic: Light Concept: We see things because light travels to the object and reflect to our eyes from object.	11
10	Topic: Light Concept: The light can exist in the area in between light source and surface of bright areas	12
11	Topic: Kinetic energy Concept: Kinetic energy is associated with the speed and the mass of an object	1, 2
12	Topic: Thermodynamics – Thermal energy Concept: Thermal energy is associated with the temperature and the mass of an object and the material of which the object is made.	3

13	Topic: Cells Concept: Both plant and animal cells perform basic life functions such as making molecules for growth.	13
14	Topic: Cells Concept: Muscle cells obtain energy from food, and they make molecules for growth.	14
15	Topic: Breathing Concept: Exhaled and inhale air consist of oxygen, carbon dioxide, nitrogen, others.	15
16	Topic: Breathing Concept: Most oxygen molecules move from the lungs to the blood by entering capillaries.	16
17	Topic: Microbes and disease Concept: Bacteria do not have intestines and lungs.	17
18	Topic: Microbes and disease Concept: Antibiotics do not work on both bacteria and viruses.	18
19	Topic: Human Body Systems Concept: Digestion is needed to break down both fat molecules and complex carbohydrate molecules into molecules that are small enough to get to cells of the body.	19
20	Topic: Human Body Systems Concept: Molecules from food and molecules of oxygen are carried by a network of arteries, veins, and microscopically small blood vessels (capillaries) to the rest of the body.	20
21	Topic: Feeding relationships Concept: The arrow in a food chain does not means ‘eats’.	21
22	Topic: Feeding relationships Concept: Food chains related by population numbers	22
23	Topic: Substances and Chemical Reactions Concept: The number of each kind of atom stays the same during a chemical reaction.	23
24	Topic: Substances and Chemical Reactions Concept: During a chemical reaction, atoms stay the same but rearrange to form new molecules.	24
25	Topic: Chemical compound Concept: Chemical compounds are pure chemicals consisting of two or several elements or atom	25
26	Topic: Chemical compound Concept: Chemical compounds are pure chemicals consisting of two or several elements or atom	26
27	Topic: Chemical equilibrium. Concept: Chemical equilibrium is influenced by volume, temperature, concentration, and pressure.	27
28	Topic: Chemical equilibrium. Concept: The catalyst speeds up the reaction but does not change the direction of the reaction	28
29	Topic: Hydrocarbons Concept: The catalyst speeds up the reaction but does not change the direction of the reaction	29
30	Topic: Hydrocarbons Concept: Tertiary carbon atoms are carbon atoms that bind to three other carbon atoms	30
31	Topic: Redox Reaction Concept: Oxidation numbers are defined as the number of negative and positive charges in an atom, which indirectly indicate the number of electrons that have been received or delivered.	31
32	Topic: Redox Reaction Concept: An oxidation-reduction (redox) reaction is a type of chemical reaction that involves a transfer of electrons between two species.	32

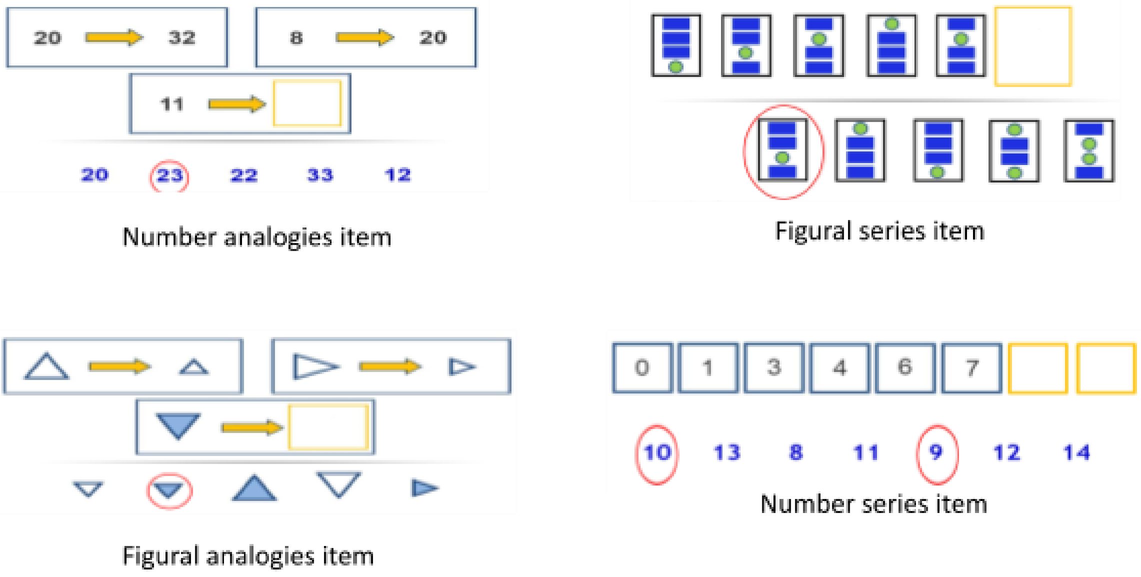
Table 11. A sample item in English Version.

Version	Sample item
English Version	1. A child has two Helium gas balloons. Balloon 1 and Balloon 2 have the same number of helium atoms.  If the thermal energy of helium in Balloon 1 is increased so that Balloon 1 has more thermal energy than helium in Balloon 2. Which helium atom will move faster than average? a) The helium atoms in Balloon 2 will be moving faster on average. b) The helium atoms in Balloon 1 will be moving faster on average. c) The helium atoms in Balloon 1 will be moving at the same average speed as the helium atoms in Balloon 2. d) The only way to tell which helium atoms will be moving faster on average is to also know the temperature of the helium in each balloon. Which one of the following is the reason for your answer to the previous question? a) Thermal energy is not related to the speed of the molecules that make up an object. b) Thermal energy is related to the speed of the molecules that make up an object. c) The thermal energy of an object is related to the gases type of the object. d) The amount of thermal energy an object has decreases as the speed of the object increases. e) ....

3.2.3 The inductive reasoning test

The inductive reasoning test was adapted and employed in this study from original version (Csapó, 1997; Pásztor, 2016). The original inductive reasoning test comprised four subscales in Hungarian and English. The inductive reasoning test has been employed in various empirical studies with different cultural contexts and school-aged samples to establish its reliability and predictive validity. These various cultural contexts include Hungary (Csapó, 1997; Pásztor et al., 2018), Finland (Hotulainen et al., 2018), Namibia (Kambeyo & Wu, 2018), Vietnam (Van Vo & Csapó, 2020), and China (Wu & Molnár, 2018). The adapted inductive reasoning test was translated into

Indonesian by two language specialists and comprises two main sections using back-and-forth translation approaches (See Appendix 5). The first section encompasses a background questionnaire on gender, grade, parents' employment, parents' education, and science and mathematics achievement scores from the previous semester. Only information related to gender and grade was used in this study. The second section included inductive reasoning items in four tasks: number analogies (NA), number series (NS), figural series (FS), and figural analogies (FA). Each subscale comprises 10 items. While a correct answer is allocated one point, incorrect answers are not awarded any points. Thus, respondents who answered all the answers correctly score a maximum of 40 points. The responses of the participants were included in the dataset automatically and in a traditional way into the Statistical Package for Social Sciences (SPSS) dataset before Rasch analysis was performed. Examples of the items in the four tasks are depicted in Figure 5.



Number analogies item

Figural series item

Figural analogies item

Number series item

Figure 5. Sample items of the inductive reasoning test based on four tasks.

3.3 Ethical consideration

The researcher obtained ethical clearance from the University of Szeged's IRB and requested a permission from the headmasters and teachers at the school (See Appendix 2). To guarantee that all ethical standards were followed, teachers and data collectors received training on research ethics. Before data gathering, participants had to fill out a written consent form (See Appendix 1). The researcher took precautions to ensure that the subjects weren't in any danger that would negatively affect them physically, mentally, economically, or socially. Participants were also selected at random using stratified random sampling to guarantee that everyone had an identical chance of participating in the study. Before starting any activity, participants were given thorough explanations of the study's nature, aim, methodology, and benefits in a language they could comprehend. The research is voluntary, and subjects are free to leave at any time during the study. In order to secure the participant information, participants' name, and related identity were recoded.

3.4 Data analysis

3.4.1 Rasch measurement

Rasch measurement is a measurement model developed by George Rasch, a Danish mathematician. Rasch measurement is based on interactions between item-person interaction and probability estimates. The interaction between items and persons can be described based on mathematical equations. Persons who have high abilities should correctly answer items with easier difficulty levels (Andrich, 2018). the probability in the measurement is governed by the difficulty of the item and person simultaneously. In other words, the probability is closely related to differences between item difficulty and individual abilities (Boone et al., 2016). Person ability and item difficulty in Rasch measurement is set based on an interval scale called logit, and item and person parameters are entirely independent (Bond & Fox, 2007; Sumintono & Widhiarso, 2014). it means that the students' ability in the measurement remains the same regardless of the item's difficulty level, and the item difficulty level remains invariant regardless of the student's ability or test takers. In this dissertation, the Rasch dichotomous model was used to analyze the two-tier multiple-choice diagnostic test, where 1 represents the correct concept, and 0 represents the misconception. The two-tier multiple-choice diagnostic test result was recorded and combined by the following procedure, in which correct responses for both items scored as 1, and incorrect response for any tier scored as 0. Unidimensionality and local independence are the two assumptions underlying Rasch measurement and the developed. The instrument must meet these two assumptions to achieve a suitable model in terms of data fit criteria. Unidimensionality is the central assumption in the single Rasch model, which shows that the items in used instruments measure the same aspect. Local independence shows the correlation between item responses, which is the latent trait of the students measured. The non-statistically significant correlation between the items used to estimate latent traits should be achieved when latent traits are controlled (Liu, 2007). The presumption of local independence prevents item redundancy and individual reliability inflation (Boone et al., 2016).

Rasch analysis was employed in this study to tackle some limitations of Classical Test Theory (CTT). The CTT has four limitations in describing a measurement model: (a) the measurement is constructed by using the result of ordinal data rather than interval scale (logit); (b) item and person in measurement are dependent; (c) measurement properties in the instrument in terms of reliability and validity are highly dependent on the sample; (d) the data is centered on group-centered statistics but is not suitable for explaining the measurement of individual respondents (Barbic & Cano, 2016).

Rasch measurement is formed on the basis of item–person interactions and probability estimates. Using equations, the interaction between the item and person can be elucidated and described. People who have low ability should not de facto be able to answer items that have a high difficulty level (Andrich, 2018). The dara are generated and determined based on a log odds unit scale (logits) as interval data, thereby ensuring that person and item parameters are entirely independent (Bond et al., 2020; Sumintono & Widhiarso, 2014). In other words, a person's ability in a measurement remains the same regardless of the item difficulty level, and the item difficulty level does not change regardless of the person's ability. For dichotomous model, the mathematical derivation of the Rasch analysis is:

$$log\frac{P_{ni1}}{P_{ni0}} = B_n - D_i$$

where

P_{ni1} or P_{ni0} is the probability that person n encountering item i is observed in category 1 or 0,

B_n is the "ability" (theta) measure of person n,

D_i is the "difficulty" (delta) measure of item i, the point where the highest and lowest categories of the item are equally probable.

(Linacre, 2021a)

In present study, WINSTEPS version 5.1.4 software (Linacre, 2021b) for Rasch measurement was utilized to perform data analysis. Rasch analysis included conducting

Rasch modelling using joint maximum likelihood estimation (JMLE) in which student scores were converted into the logit scale (interval data), ranging from negative infinity to positive infinity. Rasch parameter evaluation was employed to assess the validity and reliability based on unidimensionality, local independence, and by checking person and item reliability criteria. The Wright map was presented to confirm targeting criteria between item and person. DIF analysis was used to evaluate item bias in accordance with the test method. Rasch dichotomous model (see Section 2.6) was used to analyze the two-tier multiple-choice diagnostic test and inductive reasoning test, where 1 represents the correct answer, and 0 represents the misconception or incorrect answer.

3.4.2 Factor analysis

Factor analysis is a statistical technique for locating and examining the fundamental factors or latent variables that account for the correlations between a collection of observed variables. According to Kline (2015), factor analysis is a well-liked multivariate analysis technique that helps researchers in minimizing the complexity of data by finding the common factors that account for the variations in the data that are observed. To investigate and clarify complex data structures, the method is extensively used in the social sciences, psychology, marketing research, and other disciplines. There are two kinds of factor analysis: confirmatory factor analysis (CFA) and exploratory factor analysis (EFA) (CFA). When the researcher is unsure of the scope or makeup of the underlying variables, EFA is used. EFA reduces the number of variables by grouping them into related factors and assists the researcher in determining the factors that best describe the observed data. Contrarily, CFA is used when the researcher already knows how many and what kind of underlying variables there are. CFA determines whether the proposed factor structure corresponds to the data collected (Brown, 2015). In conclusion, factor analysis is a statistical method for investigating and explaining complicated data structures by locating the underlying factors that contribute to the observed correlations between variables. The two major varieties of factor analysis are EFA and CFA, and each has particular uses and benefits. The research question, the data structure, the researcher's previous knowledge and presumptions regarding the underlying factors, and the chosen factor analysis method are all factors.

In this dissertation, CFA was used to checking construct validity. To perform CFA, we employed MPLUS 8.4 (Muthén & Muthén, 2017) with two CFA models with the ULS estimator, as it provides more accurate results regarding standard errors, estimates and fit indices than weight least square (WLS) or maximum likelihood (ML) (Muthén 1993). CFA evaluated the model based on standardized root mean square residual (SRMR), comparative fit index (CFI) and the root mean square error of approximation (RMSEA). Goodness-of-fit indices measured how well the rotated matrix matched the original matrix. CFI required a large number of values and compared the real correlation matrix with the reproduced correlation matrix. RMSEA and SRMR pertain to the value of residual statistics, which are expected to be small in the residual matrix. Hence, we observed the following cut-off values to assess model fit: SRMR < .08, CFI > .90 and RMSEA < .06 (Caleon & Subramaniam, 2010; Hu & Bentler, 1999).

3.4.3 The descriptive and Inferential and analysis

Descriptive analysis is a statistical technique used to enumerate and describe the features of a data collection, such as mean, standard deviation, and frequency distribution. In addition to describing the data in a straightforward and concise manner, it also helps to spot patterns and trends. On the other hand, inferential analysis utilizes a sample to draw conclusions or generalizations about a community. In this process, theories are tested, parameters are estimated, and the strength of the correlation between variables is measured (Gall et al., 2007). The use of analysis of variance is a typical method for applying inferential analysis to educational study (ANOVA). ANOVA is a statistical technique used to compare the means of two or more groups to see if there are statistically meaningful differences between the groups. For instance, a researcher may use ANOVA to compare the academic performance of students in three different classes to see if there are any significant variations in performance between the classes (Tabachnick & Fidell, 2019). In this dissertation, all empirical studies was conducted involving descriptive and inferential analysis by utilizing The SPSS version 25 (IBM Corp, 2017).

3.5 Cross sectional study

Cross-sectional studies are a type of research design that is employed in academic studies to collect data at one moment in time and examine the relationships between various variables (Creswell & Creswell, 2017; Leedy & Ormrod, 2005). The design is employed to spot patterns or trends in data or to try theories regarding the frequency of particular traits in a population (Merriam & Tisdell, 2015). Examining various aspects of misconceptions in science, such as such as exploring students' understanding, measuring item difficulty level and assessing inductive reasoning skills, can be done using cross-sectional studies (Park & Liu, 2019; Pásztor, 2016; Van Vo & Csapó, 2023).

A longitudinal study design is a type of research methodology that includes keeping track of a group of people over time and gathering data at various points during that time (Cohen et al., 2018; Creswell & Creswell, 2017). With the help of this design, researchers can track changes in the study subjects over time, which can reveal important details about the evolution of various phenomena like learning results, attitudes, and behaviors. A study by Seo et al. (2019) found that the ability to examine how various elements, such as belief and mathematics ability, contribute to changes in student STEM achievement over time makes longitudinal study design particularly helpful in educational research. Furthermore, longitudinal research can be used to pinpoint possible causal links between various educational initiatives.

In this dissertation, researchers gather information from a representative sample of participants during a cross-sectional study at a particular moment. The sample is chosen to guarantee that it accurately reflects the traits of the target community (Yin, 2018). To make sure the sample is representative, different sampling methods can be used, such as stratified sampling or random sampling (Creswell & Creswell, 2017; Leedy & Ormrod, 2005). The sample number should be sufficient to guarantee statistical power and the validity of the findings (Merriam & Tisdell, 2015).

Cross-sectional studies have the advantage of being relatively simple and affordable to perform because participants do not need to be followed up with over an extended period of time (Creswell & Creswell, 2017). The design, however, has drawbacks, such as the inability to prove causation and the possibility of confounding variables (Bryman, 2016; Cohen et al., 2018; Merriam & Tisdell, 2015). To make sure that the findings of a cross-sectional study design are accurate and reliable, researchers must carefully take into account these limitations (Merriam & Tisdell, 2015; Yin, 2018). In this dissertation, cross-sectional studies are used to investigate various factors to investigate misconceptions in science and inductive reasoning, such as evaluating misconceptions in sciences based on different groups, evaluating item difficulty and student interaction, and exploring student misconception based on different science subjects. Table 12 illustrates the cross-sectional studies that had been conducted in this dissertation from two different datasets in pilot and main study.

Table 12. Cross-sectional studies from pilot and main study in this dissertation.

Timeline	Main objective	Instrument	Sample
May to June 2019 (pilot study)	1. Conducting pilot study 2. Checking the psychometric properties of the developed instrument 3. Examining student misconceptions in science learning 4. identifying background factors affecting student misconceptions in the learning context.	1. Background questionnaire 2. The two-tier multiple-choice test	10th, 11th, and 12th N =152
September – June 2021 (main study)	1. Investigating item difficulty patterns 2. Evaluating item–person map interaction 3. Checking the DIF based on gender and grade across science disciplines	1. Background questionnaire 2. The two-tier multiple-choice test 3. IR Test	10th, 11th, 12th and PST N =856
September – June 2021 (main study)	1. Investigating student misconceptions in science concepts across school grades 2. examining student–item interaction regarding science concepts 3. detecting outliers in student misconceptions 4. predicting background factors that influence students' misconception in sciences	1. Background questionnaire 2. The two-tier multiple-choice test 3. IR Test	10th, 11th, 12th and PST N =856
September – June 2021 (main study)	1. Assessing the adapted Indonesian version of the inductive reasoning test 2. Classifying their inductive reasoning levels in accordance with grade and gender.	1. Background questionnaire 2. The two-tier multiple-choice test 3. IR Test	10th, 11th, 12th and PST N =856

CHAPTER 4. EMPERICAL STUDIES

4.1 Evaluation and development of students' misconceptions using diagnostic assessment in science across school grades: A Rasch measurement approach

4.1.1 Introduction

The literature review investigations were carried out to find out various concepts in learning science that distribute misconceptions to students and instruments that can identify misconceptions. The preminarily systematic literature review by Soeharto et al (2019) has found 111 articles from 2015 to 2019 have focused on student misconceptions in science. In this study, sixteen concepts in science subjects are selected based on by Soeharto et al (2019). Soeharto et al (2019) found that in the development trend of using diagnostic tools to identify misconceptions, the multiple-tier test (33.06%) is the most diagnostic tool used to identify science misconceptions. Therefore, we decided to develop the two-tier multiple-choice test assisted with the Rasch measurement model and to identify and evaluate the development student misconception across school grade and gender. We performed Rasch measurement model because Rasch measurement can convert research instruments like Physics measurement tools having interval scale. Rasch measurement also can tackle weakness of CTT analysis from previous studies (e.g., (Galvin & Mooney, 2015; Laliyo et al., 2019; Taslidere, 2016). Therefore, this study was focused on developing diagnostic assessment to measure and investigate student misconceptions in science.

4.1.2 Method

The quantitative approach was employed, where a two-tier test multiple-choice test was administered to understand student misconceptions in science, especially in physics, biology, and chemistry, and Rasch modelling was used to analyze psychometric properties.

Participants

The participants in this preliminary study were 153 students at public senior high schools and private senior high school schools in Pontianak, part of West Kalimantan province, Indonesia. The samples were drawn using stratified random sampling according to student grades. In this study, we selected five classes from 5 different schools randomly to be analyzed. Data were collected from 123 students using the paper-based test and 30 students using the online Electronic Diagnostic Assessment System, the eDia, developed by the Center for Research on Learning and Instruction at the University of Szeged (Csapó & Molnár 2019). The eDia system can support item writing, editing, and scoring using logfile analysis as well as administering the test, and giving feedback. The eDia was used in the various research areas in teaching and learning, including reading, science, and mathematics, that can be accessed using internet browser applications such as Google Chrome and Firefox (Csapó & Molnár 2019; Greiff et al., 2018). The demographic profile of the participants is presented in Table 13. This study was conducted from Mei to June 2019. Students spent 120 minutes completing the test under the surveillance of researchers and teachers.

Table 13. The demographic profile of the participants in this study (N=153).

Demographic		Frequency	Percentage (%)
Gender	Girls	68	44.4
	Boys	85	55.6
Grade	10th	57	37.3
	11th	55	35.9
	12th	41	26.8
School category	Public	109	71.2
	Private	44	28.8
Living place	City	77	50.3
	District	76	49.7

Instruments

The background questionnaire was adapted from the Indonesian version of the PISA 2015 SES instrument (OECD, 2016) (See Section 3.2.1). To capture student misconceptions or alternative conceptions, we implemented the developed two-tier multiple-choice diagnostic test (See Section 3.2.1). The two-tier test cannot differentiate students who are just guessing answers and related confidence level, and some researchers usually applied CTT analysis and the CRI (Hasan et al., 1999). Otherwise, Rasch measurement can overcome the weakness of two-tier tests with CTT and CRI analysis in cases of the certainty level and can provide a comprehensive and objective measure (Barbic & Cano, 2016). Before constructing and developing the instrument, the researcher investigated some literature review studies and misconceptions in science handbooks (AAAS, 2019; Allen, 2014; Csapó, 1998; Soeharto et al., 2019). This process was conducted to find common rationales behind misconceptions in science. Sixteen concepts were selected and adjusted to the Indonesian education curriculum for

Curriculum 2013, especially on the senior high school level from the physics, biology, and chemistry concepts represented in Table 2. Thirty-two item questions were adapted developed in the form of a two-tier multiple-choice diagnostic test with eight items is adapted from the American Association for the Advancement of Science (AAAS), two items adapted from (Csapó, 1998), 23 items newly designed by authors. The backward–forward translation process from English to Indonesian was conducted by two science and mathematics instructors and researchers. Table 14 summarize Concepts and item number in the developed two-tier multiple-choice diagnostic test.

Table 14. Concepts and item number in the developed two-tier multiple-choice diagnostic test (N=32).

Subject	Concept	Item numbers	Total item
Physics	Kinetic energy, thermodynamics–thermal energy, atoms and molecules, impulse and momentums, light, and force	1, 2, 3, 4, 5, 6, 7, 8, 9,10, 11, 12	12
Biology	Human body systems, cells, breathing, feeding relationships, microbes, and disease	13, 14, 15, 16, 17, 18, 19, 20, 21, 22	10
Chemistry	Chemical compounds, substances and chemical reactions, redox reaction, hydrocarbons, and chemicals equilibrium	23, 24, 25, 26, 27, 28, 29, 30, 31, 32	10

The two-tier multiple-choice diagnostic test consists of two-level questions. The first-tier question asks about science content, and the second-tier question asks about scientific reasoning. Students can choose one choice in the second tier or write down their own reason in the form of an open-ended answer to explain the related science content. Peterson et al. (1989) supported this two-tier test format since most multiple-choice questions did not provide sufficient information to explain the students' reasoning, whereas the additional explanation items in second-tier questions can assess students' understanding related to science concepts and diagnose misconceptions.

Procedures, data analysis, and Rasch measurement

Before conducting data collection in schools, researchers asked permission to administer the tests at schools and granted ethical research approval. The paper-based tests were conducted in student classrooms with the guidance and supervision of researchers and teachers. online tests were conducted in each school computer laboratory using the eDia system. Statistical Package for the Social Sciences (SPSS) version 25 (IBM Corp, 2017) and the WINSTEPS version 4.7.0 software (Linacre, 2020) were employed in this study. WINSTEPS version 4.7.0 software was used to perform data analysis using Rasch modelling. WINSTEPS software performed Rasch analysis from simple rectangular dataset. WINSTEPS can be utilized to analyze multiple-choice, dichotomous, and multiple rating-scale and partial credit items. This software can be downloaded in trial as a full version in WINSTEPS website, www.WINSTEPS.com, and the SPSS version 25 was used to analyze using statistical methods such as descriptive statistics, regressions, and ANOVA. All samples in the data set were investigated because this preliminary study wanted to explore item and person interaction.

The most common method is used to analyze an instrument's psychometric quality by employing software or statistic calculations based on Classical Test Theory (CTT) principles. However, CTT has several limitations, such as the sample-dependent and biased derived scores against central scores (Bradley et al., 2015). in CTT, missing data presents a problem in calculating the data. Reliability measures are described using Cronbach's alpha, and measurement evidence is based on the correlation between items and other measures, which may not be reliable and valid. It is challenging to assess individual items' characteristics to determine the effectiveness of items in the population and their contribution to measuring the overall latent construct. There are many measurement problems with surveys, questionnaires, and rating scales, which concluded that measurement using CTT could produce various responses and analysis biases (Bradley et al., 2015; Zlatkin-Troitschanskaia et al., 2015). Therefore, the Rasch measurement was employed to tackle measurement issues in CTT (Barbic & Cano, 2016). Rasch analysis can explain the difficulty level of an item accurately and precisely, detect the suitability and interaction of items and persons (item-person maps), identify outliers (person misfit), detect item bias (differential item functioning (DIF)), which is useful for describing and identifying students conceptions in sciences in this study (Boone et al., 2016; Sumintono & Widhiarso, 2014b).

4.1.3 Results

Scaling, reliability, and validity of the developed instrument

We analyzed the psychometric properties of the developed instrument based on Rasch measurement model. WINSTEPS run the analysis based on the Joint Maximum Likelihood Estimation (JMLE) equations; in this formulation, the raw data were converted to interval data (logit) (Linacre, 1998, 2020). The logit scale can express person ability and item difficulty ranging from positive infinity to negative infinity. The 32 items of the misconception test and 153 participants were processed with a two-facet item and person model using the Rasch measurement model with the WINSTEPS software. The mean measure (logit) of the items is 0.00, and the standard deviation (SD) is relatively high (1.84), which means that the variation or dispersion of item measurement in terms of item difficulty was wide across the logit scale. The mean measure was 0.75 logit for students, indicating all respondents tended to be strongly involved in misconception in science, but the person SD was 0.87, almost achieving 1, showing person variation is ideal for data analysis. The mean OUTFIT mean-square and The average outfit z-standardized (ZSTD) was acceptable (ranging from -2 to +2), and outfit mean-square (MNSQ) statistics are 0.96, which is near their expected value of 1 for item and student, and the chi-squared score showing the data achieve the normal distribution criteria and Rasch model fits globally (Boone et al., 2013; Engelhard Jr, 2013; Linacre, 2020). The item separation was 5.81, indicating various levels of item difficulties, and the person separation was 1.91 showing that data consists of 2 levels, high and low performance. The reliability of items and person were excellent (Fisher, 2007; Taber, 2018). The summary statistics of item and person can be seen in Table 15.

Table 15. The summary statistic based on persons and items.

	Persons	Item
N	153	32
Measure	0.75	0
Mean	19.7	94
SD	0.87	1.84
SE	0.08	0.33
Mean Outfit MNSQ	0.96	0.96
Mean Outfit ZSTD	0.12	-0.09
Separation	1.91	5.81
Reliability	0.76	0.97
Cronbach's Alpha	0.8	
Chi-squared (χ2)	4443.85 (df= 4431)	
Probability	0.4429*	

*Normally distributed

The reliability is calculated based on item internal consistency using Cronbach's alpha value for all items and based on the item and person reliability parameter in Rasch measurement. Cronbach alpha for the whole item was 0.8 that indicated high internal consistency reliability (Taber, 2018). The reliability parameter in Rasch measurement was 0.76 and 0.97 for person and item statistics representing good reliability (more than 0.67) (Fisher, 2007). all items in the developed instrument are not deleted and retained in the developed instrument. To achieve validity, we assessed the unidimensionality and local independence of the instrument. The unidimensionality shows that the instrument measures the same dimension, which is student misconception in science. The instrument can achieve unidimensionality if the value of the raw variance explained by the measure is more than 30% (Chou & Wang, 2010; Linacre, 1998). The analysis result confirmed that the developed instrument passed the minimum threshold for the variance explained by measure was 37.4% with 12.18 eigenvalue. The local independence is achieved when the raw residual correlation between items is lower than 0.3 (Christensen et al., 2017; Hagell, 2014). The instrument's local independence in this study was below 0.3, which indicated that no items have local dependence. The test information function in Figure 6 have given additional proof of test quality to measure student misconception in science with large range of difficulty level from -8 to +8. It means that the develop test can cover from the easiest difficulty item to the most difficult item base on person ability. Therefore, we can conclude that the developed two-tier multiple-choice used in this study is valid and reliable.

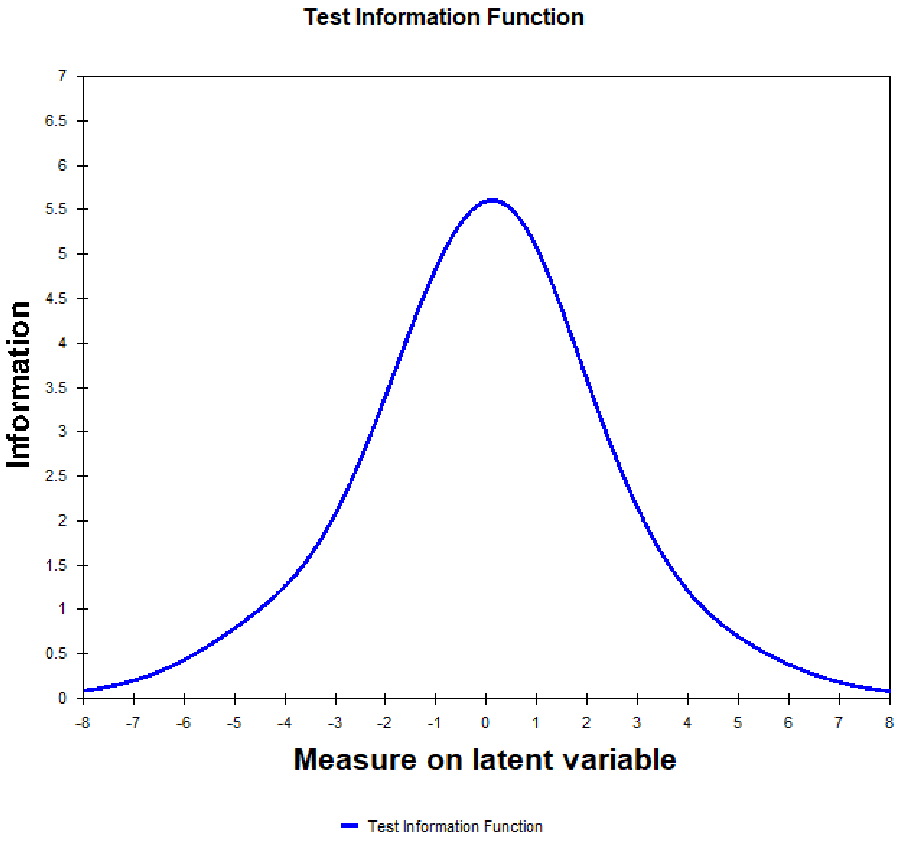


Figure 6. Test Information Function for the two-tier multiple-choice test.

Item fit

Item fit analysis was carried out to see whether the developed two-tier multiple-choice diagnostic test could measure student misconceptions at the senior high school level. The ideal MNSQ outfit and infit value are 1 based on the Rasch measurement model, but the acceptable values ranging from (0.5-1.5) below 1.6 are still acceptable, and besides that it can also be seen based on the point measure correlation range from 0.4 to 0.85 as an additional indicator (Andrich, 2018; Bond & Fox, 2007). The results of the analysis show that the mean of infit and outfit MNSQ is 0.99 (SD = 0.18) and 0.9 (SD = 0.39), respectively. However, there are 3 misfit items based on the MNSQ outfit value, namely items PHY1 (0.11), PHY3 (0.23), and CHEM32 (0.36). These three items must be removed or corrected first before administering the test in larger sample. The item measure is calculated in logit units ranging from the least difficult (-4.86 logit) to the most difficult (5.05 logit), which means that the instrument is around 4 or 5 categories in the item difficulty level. However, because this study is the preliminary study for developing instruments, those three items are retained to item analysis and improvement other test version in future study. The distribution of item fit order is shown in Figure 7.

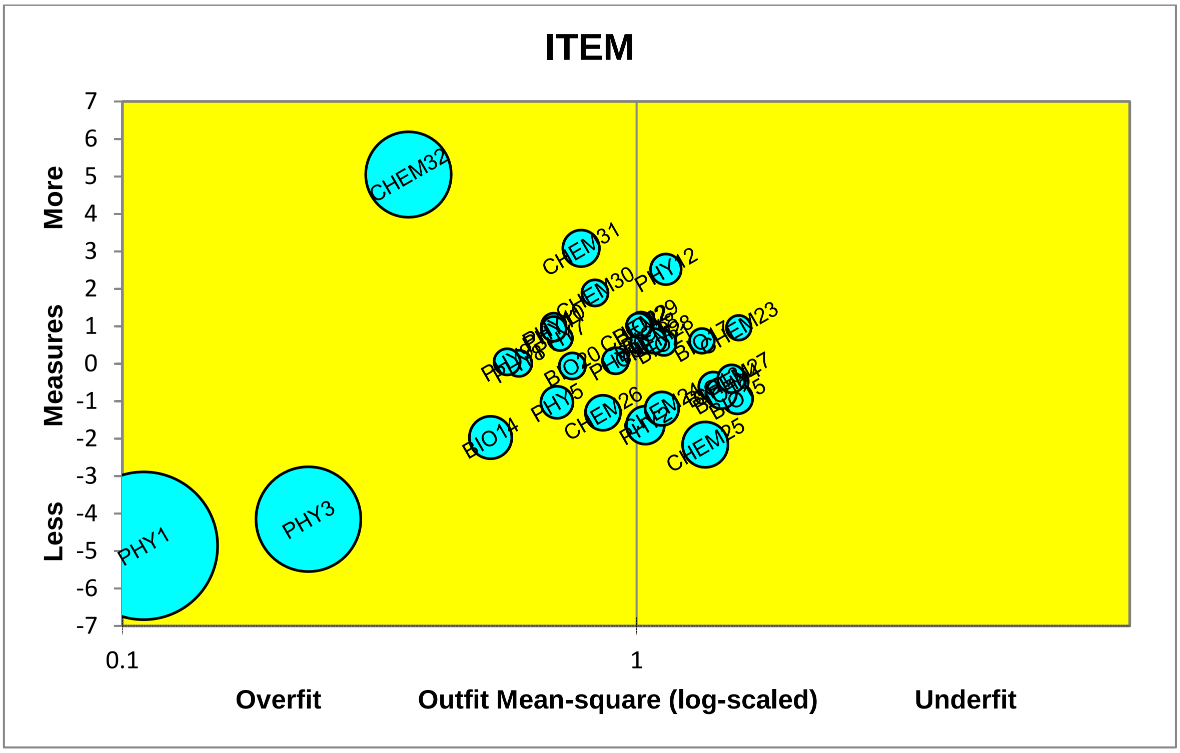


Figure 7. Buble chart of item fit order based on infit MNSQ.

Person ability

Person ability measure describes the student ability in answering items on the test. Person ability in this study ranging from -2.11 logit to 2.43 (M = 0.75, SD = 1). We categorized person ability into 4 types on logit value of item (LVI) based on Sumintono & Widhiarso (2014), low misconception 16.33% (2.43 <LVI <1.75), moderate misconception 49.01% (0.75 <LVI <1.75), high misconception 14.37% (0.75 < LVI <- 0.25), and very high misconception 20.26% (-0.25 <LVI <- 2.11). Overall, 37% of students answered incorrectly, which shows that students have misconceptions on the basic concepts in science learning. Misconceptions in each subject in science were also checked based on the percentage of students' incorrect answers to see how the misconceptions were distributed based on the science subjects, physics (33.4%), biology (35.22%), and chemistry (47.97%).

```
MEASURE                                PERSON - MAP - ITEM
                                            <more>|<rare>
     5                                            +  CHEM32
                                                  |
                                                  |
                                                  |
                                                  |
                                                  |
     4                                            +
                                                  |
                                                  |T
                                                  |
                                                  |
     3                                            +  CHEM31
                                                  |
                                                 T|
                                                  |
                                   P P P P P P P  |  PHY12
                                                  |
     2                                 P P P P P  +
                       P P P P P P P P P P P P P  |S CHEM30
                                                 S|
                             P P P P P P P P P P  |
                         P P P P P P P P P P P P  |
                 P P P P P P P P P P P P P P P P  |
     1 P P P P P P P P P P P P P P P P P P P P P  +  BIO22  CHEM23 CHEM29 PHY10
                                                  |  BIO18  PHY11
                 P P P P P P P P P P P P P P P P M|  BIO17  BIO19  CHEM28 PHY7
                             P P P P P P P P P P  |  BIO21
                                               P  |
     0                                     P P P  +  PHY6
                                                  |M BIO20  PHY8   PHY9
                                 P P P P P P P P S|
                                       P P P P P  |  CHEM27
                               P P P P P P P P P  |  BIO13  PHY4
                                   P P P P P P P  |  BIO16
    -1                                       P P  +  BIO15
                                           P P P  |  PHY5
                                               P T|  CHEM24 CHEM26
                                                  |
                                                  |  PHY2
    -2                                         P  +S
                                                  |  BIO14
                                           P P P  |  CHEM25
                                                  |
                                                  |
                                                  |
    -3                                            +
                                                  |
                                                  |
                                                  |
                                                  |
                                                  |T
    -4                                            +
                                                  |  PHY3
                                                  |
                                                  |
                                                  |
    -5                                            +  PHY1
                                            <less>|<freq>
```

Figure 8. The Wright item-person map of student misconception in science subjects.

To comprehend the interaction between item and person, we ran the item-person analysis using the Wright map, illustrating the student ability on the left side and item difficulty on the right side. The Wright map is item-person maps that can compare items and people simultaneously in the context of a measurement on the one interval scale (logit), and assess the interactions between items and person, as well as check students' individual abilities. If the item is in line with the person, it means that the student has a

50% chance (p = 0.5) of answering correctly because the difficulty level of the item is the same as student ability. If the person is located above the item, it means that the student has the correct chance to answer the question more than 50% (p> 0.5), and if the difficulty level of the item is higher than the student's ability, the chance of the student to answer correctly will decrease from 50% (p <0.5) (Griffin, 2010; Linacre, 2020). In this study the easiest item is shown at the bottom on the right of y axis (CHEM25, PHY1, and PHY3). and the most difficult item is shown at the bottom on the right of y axis (CHEM31 and CHEM32). The good items in the instrument have to cover all student abilities in the item-person map (Griffin, 2010). However, there are three misfit items, which are CHEM32 (too difficult), and PHY1 and PHY3 (too easy) having logit more than two standard deviation. In general, if we omit misfit items, the test still shows good performance and acceptable because the developed test can cover all scales of person abilities. Therefore, we can conclude that the developed test is matching with the target group of in testing student misconception in science subjects. The Wright item-person map of student misconception in science subjects can be seen on Figure 8.

Item bias based on Differential Item Functioning (DIF)

DIF analysis was conducted to check whether there were items bias based on gender. DIF analysis suggested on participant responses based on subgroups for each item in the test of measuring student misconceptions on science learning (Adams et al., 2020; Boone et al., 2014; Rouquette et al., 2019). DIF analysis is divided into three types namely negligible, slight to moderate (| DIF | ≥ 0.43 logits), moderate to large (| DIF | ≥ 0.64 logits) (Zwick et al., 1999). DIF analysis (Figure 8) shows that the items PHY1 and CHEM32 have DIF bias in the moderate to large category. These two items was also misfit item. Items PHY1 and CHEM32 explained that these two items were more difficult for boys than girls to answer correctly.

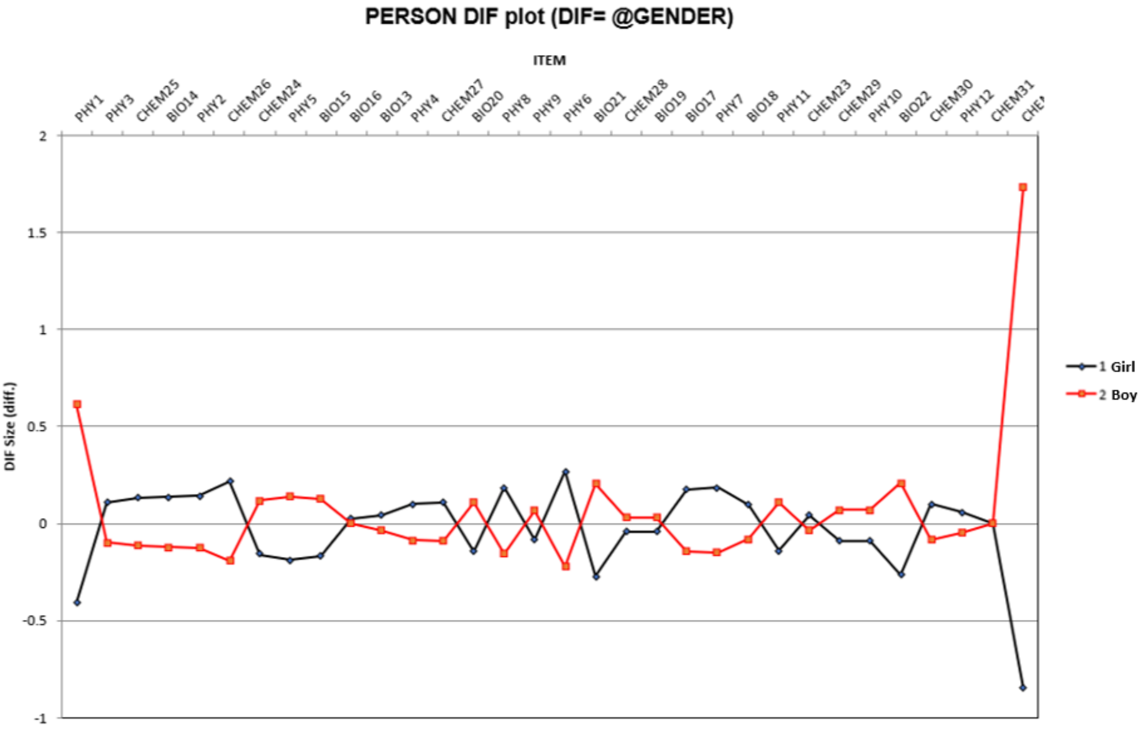


Figure 9. DIF based on gender

The differences of student misconceptions in science-based school grade

ANOVA was conducted to determine the comparison of student misconceptions across school grades of student misconceptions in science on the test and subtest. The analysis showed that there were significant differences between school grades which confirmed student misconception test and subtest score across four cohorts with the Physics subtest [F (2, 152) = 6.35, p <.01], Biology subtest [F (2, 152) = 7.84, p <.01], Chemistry subtest [F ((2, 152) = 5.06, p <.01], The entire test [F (2, 152) = 10.93, p <.01]. Because the equal variances are not assumed, we ran Dunnett T3-test to identify specific differences between the school grades in Table 16. Dunnett T3-test was utilized when comparing one group to other groups. Dunnett T3-test is the most powerful ANOVA post-hoc tests than others. Overall, the entire test's significant differences were found for all school grade pairs, except for the differences in all subtests (Physics, Biology, and Chemistry), which showed that the 10th-grade students had a higher mean score of misconceptions than the 11th-grade students on the subtest and the entire test. This trend showed that 10th-grade students misunderstand science concepts more than 11th-grade students. However, the 12th-grade students suffered the highest conceptual misconceptions than students in the 10th- and 11th-grades. This phenomenon showed misconceptions that are resistant, persistent to change, and rooted deeply in science concepts made students at a higher level more difficult to understand science learning.

Table 16. The Dunnett-T3 multiple comparisons of student misconception on school grades (N=153).

Grade	Physics		Biology		Chemistry		Test	
	Mean differences	p	Mean differences	p	Mean differences	p	Mean differences	p
10th & 11th	0.58	.54	0.06	.99	0.30	.72	0.93	.56
10th & 12th	-1.35	.06	-1.31*	.01	-0.82	.07	-3.61*	.00
11th & 12th	-1.94*	.00	-1.38*	.00	1.13*	.01	-4.55*	.00

Gender differences among school grades

In general, the boxplot showed that boys and girls were identified as having equivalent mean scores of student misconception in science for each cohort shown in Figure 17. Mean scores of student misconceptions in science range from 0.28 to 0.47, where the mean score for boys in 12th-grade (0.47) explained that boys were suffering misconceptions higher than girls (0.44). However, overall for whole grades, the average score among boys and girls is relative at the same level. The length of the boxplot in Figure 10 showed that the standard deviation for the 12th grade is higher than the 10th- and 11th-grade, showing that girls experience more varied misconceptions than boys. Table 18. Comparing boys and girls based t-test with the maximum likelihood estimate of the students' conceptual misconception in science. No significant differences were found in the test and whole grade school level (p > .05). This also indicates that each cohort is not different between girls and boys. Therefore, we can conclude that the estimate of girls and boys had an equivalent value. However, unexpectedly, in the 12th-grade boys, the mean score of student misconceptions in science was slightly higher than girls, and the opposite was the variation in misconceptions where the misconceptions of female students were more varied than boys.

Table 17. The t-test comparing student misconceptions between girls and boys (N=153).

Grade	Girl		Boy			
	N	Mean (SD)	N	Mean (SD)	t	p
10	13	0.35(0.12)	44	0.34(0.13)	0.16	.60
11	37	0.33(0.14)	18	0.28(0.09)	1.51	.11
12	18	0.44(0.21)	23	0.47(0.18)	-0.48	.41

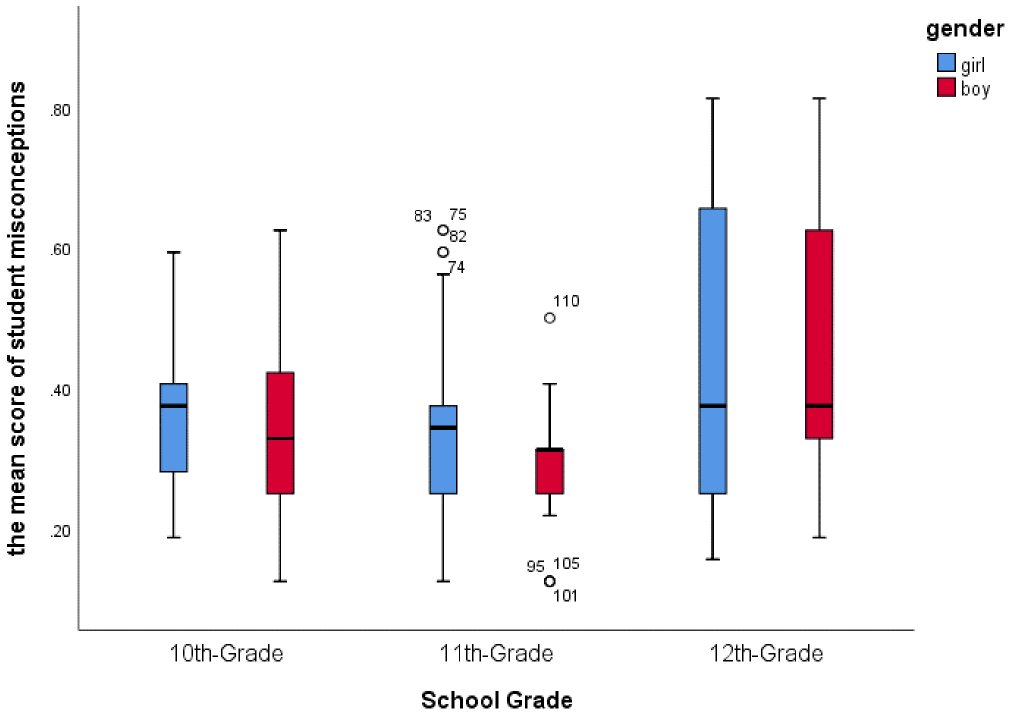


Figure 10. Comparison of student misconception among school grades.

Table 18 showed student misconceptions for all science subjects. Boys (48%) and girls (47%) suffered from high misconceptions in chemistry subject. However, overall, boys and girls had the same or equivalent percentage of misconceptions, and no significant differences were found based on the t-test conducted on all science subjects. These results were in line with the study about student misconceptions in science on gender subgroups (Taslidere, 2016; Treagust, 1988; Tsui & Treagust, 2010).

Table 16. The t-test comparing misconceptions in science subjects between girls and boys (N=153).

Subject	Girl	Boy		
	Mean (SD)	Mean (SD)	t	p
Physics	0.34 (0.23)	0.32 (0.22)	0.29	.38
Biology	0.34(0.21)	0.35(0.18)	-0.46	.15
Chemistry	0.48(0.18)	0.47(0.17)	.070	.40
All subject (Science)	0.37 (0.16)	0.36 (0.16)	0.05	.70

Predicting student misconceptions in science

To explore how other factors predict student misconceptions in science, we ran the stepwise multiple regression with school category, school grade, father education, mother education, school performance as predictors. The analysis result showed that only school grade predictor could significantly explain 25.2% of the variance on student misconception mean scores, F (152) = 10.208, p <.01. These results indicated that grade school is an essential factor in developing student misconceptions in learning science at senior high school.

4.1.4 Discussion

The preliminary result indicated that the developed two-tier multiple-choice diagnostic test is valid and reliable for identifying student misconceptions in science for grades 10th, 11th, and 12th in school contexts. All items on the test can identify conceptual misconceptions and cover all student ability areas even though three misfit items must be revised or deleted in further research. the item-person analysis indicated that all item can cover student ability from low to high ability although three misfit items have to be revised for further research in large sample. Nonetheless, the stabilized value for misfit item depend on the number of samples (Boone et al., 2013; Khine, 2020; Planinic et al., 2019). In the development of student misconceptions, we found that there are significant differences in student misconception between science disciplines based on ANOVA test. this findings are in line with previous studies in student and item evaluation related to energy (Park & Liu, 2019), which is one of science concepts chosen in this study. Student misconception mean scores in science may range across school grades but remain persistent and resistant to the same concept indicating student still suffering misconception in science even if they have been in upper grade level (Taslidere, 2016; Tsui & Treagust, 2010; Wandersee et al., 1994). Moreover, the finding in Figure 5 showed that students at 12th grade had higher misconceptions than students in grades 10th and 11th, but. The 10th and 11th-grades' pairs did not have substantial significant misconception score. This condition might occur based on characteristic of misconceptions that are persistent, resistant, and root deeply in science concepts (Arslan et al., 2012; Wandersee et al., 1994) whereby students in grade 12th actually already had misconceptions related to particular science concepts when they were in grade 10th and 11th so that student misunderstandings were getting worse by time in the upper grade level, 12th grade. The DIF analysis showed that two items are biased based on gender, PHY1, and CHEM32. However, these items are still retained to analyze the psychometric properties of the developed test.

The online and paper-based tests in this study offers several solutions to the initial stages' instrument development process. This study might be the first study that assesses student misconceptions in science based on the Rasch measurement model.

Rasch measurement can solve several problems in assessing misconceptions that cannot be resolved based on CTT, for example, detecting the difficulty level of an item accurately and precisely, determining the misfit of items and persons, and identifying DIF items (Adams et al., 2020; Boone et al., 2013). Technology-based testing offers several solutions to cover an even broader competency range in development tests on different difficulty levels. This present study identifies student misconceptions in science subjects, physics (33.4%), biology (35.22%), and chemistry (47.97. In comparing school grades, based on school grade regression analysis was able to explain 25.2% variance of student misconceptions in science. stepwise regression showed that only school grade predictor could significantly explain 25.2% of the variance on student misconception mean scores, F (152) = 10.208, p <.01.

4.1.5 Conclusion

To sum up, we conclude that this study can provide comprehensive knowledge related to evaluation and development of student misconceptions in science. All the items in the developed instrument are valid and reliable covering student ability based on item-person Wright maps although there are three misfit item and DIF issue based on gender. The ANOVA test have verified that there are significant differences between science concepts across science disciplines and school grades whereby grade school predicted student misconception in science based on stepwise multiple regression. Independent sample t-test verified that no significant difference was found between boys and girls.

There are several limitations in the measurement in this study as well. We did not develop items based on all scientific concepts studied in Indonesia. Items selected are based on concepts that distribute misconceptions in the previous research (AAAS, 2019; Allen, 2014; Csapó 1998; Gurel et al., 2015; Soeharto et al., 2019). Therefore, further research is needed to find new science concepts, where students find it challenging to understand and distribute conceptual misconceptions. The participants also were drawn from a small population in West Kalimantan province may be a limitation in this study. We realized that some of the results in the educational context could not be generalized.

This research is early-stage research, so it is necessary to research a larger sample to identify misconceptions in science at school contexts. However, this research is probably the first in using Rasch modelling analysis in developing a two-tier test by combining online and paper assessments. This study's exposure might encourage the emergence of other studies related to scientific misconceptions with Rasch measurement analysis. We hope the successfully developed instrument will inspire other researchers to create a diagnostic assessment based on Rasch measurement. For educators and instructors, we hope that our report related to evaluation and development of student misconceptions in science can be an initial signal or alert to overcome student problem in understanding science concepts. if educator realize what is the specific concepts that difficult to understand in learning activity, they can cope the problem and be more concerning to design proper and correct lesson plan to make student understand and to improve student science performance.

4.2 Evaluating item difficulty patterns for assessing student misconceptions in science across physics, chemistry, and biology concepts

4.2.1 Introduction

Science concepts are critical elements in explaining and understanding natural phenomena across all science disciplines. The particular science concept provides a practical framework for integrating science disciplines and has a significant impact on the learning process and in the thinking and modeling of natural and technological processes. Several studies reported students' difficulties in learning scientific concepts. Soeharto (2017) reported that students suffered misconceptions about impulses and momentum because of a lack of understanding of various types of collision. Additionally, Tiruneh et al. (2017) found that students experienced difficulty in solving critical thinking problems related to electricity and magnetism. Students' weaknesses in understanding science concepts across science disciplines are attributed to how some science concepts are introduced and applied in varied ways that are often incompatible (Cooper & Klymkowsky, 2013; Lancor, 2014). Understanding the science concepts properly will help students to work on problems of varying degrees of difficulty. Thus, the investigation of difficulty levels of science concepts across science disciplines has the potential to hamper students through suffering misconceptions and thereby failing to achieve their best performance in science.

Several studies have been conducted to investigate conceptions in various science concepts across various disciplines (e.g., Butler et al. (2015), Korur (2015), Park and Liu (2019), Peterson et al. (1989), Tiruneh et al. (2017), Tümay (2016)). However, comparing the actual difficulty level of science concepts from various disciplines becomes a problem and is challenging to implement. The results of an instrument test may reflect differences in the respondents' abilities and the lack of ability to work on questions in various science disciplines. Hence, there is a necessity to create an instrument that allows for a standardized measurement of science concepts in various scientific fields so that teachers can recognize the especially challenging science concepts whenever they teach students in various areas of science.

The goal of objective measurement is located at the core of science, and science education research should also attempt to carry the instrument according to objective measurement criteria. Our study evaluates item difficulty estimates using a standardized instrument to assess the distributed science concepts misconceptions to students across the science disciplines using Rasch measurement and explores DIF. Although some research concentrates on students' science conceptions of particular concepts, to what extent students have experienced the ease or difficulty in understanding science concepts has not been fully elucidated using a standardized instrument to measure concepts comprehension across science disciplines. This study will fill the gap in empirical research that provides evidence related to students' difficulties in understanding science concepts across disciplines, especially science concepts that generate misconceptions in students on the basis of key concepts in the findings of previous research findings by Soeharto et al. (2019). Previous studies on pre-service science teachers and undergraduate students are limited (Singer, 2013), and some studies focus more on students at the secondary school level (Erman, 2017; Slater et al., 2018; Tiruneh et al., 2017; Tümay, 2016). This study will target both groups, students at secondary school and teachers who have completed pre-service and are undergoing education based on the Indonesian science core curriculum.

4.2.2 Methods

Participants

Participants were 856 students from senior high school students and university students (pre-service science teachers), West Kalimantan province, Indonesia. We selected 11 classes randomly from five different schools in total as representative schools in this area. All participants in this study were students from three different school levels, 10th, 11th, and 12th grades, and pre-service science teachers. The paper-based test was administered at the schools and university. Students and pre-service science teachers spent 120 minutes completing the test under the supervision of researchers and teachers. Table 19 presents the demographic characteristics of the participants.

Table 19. Demographic characteristics of participants in this study (N=856).

Demographic characteristics		Frequency	Percentage (%)
Gender	Females	448	52.3
	Males	408	47.7
Grade	10th	231	27.0
	11th	291	34.0
	12th	153	17.9
	Pre-service science teacher (PST)	181	21.1
School category	Public	621	72.5
	Private	235	27.5
Living place	City	444	51.9
	District	412	48.1

Instrument

The background questionnaire was adapted from the Indonesian version of the PISA 2015 SES instrument (OECD, 2016) (See Section 3.2.1). To capture student misconceptions or alternative conceptions, we implemented the developed two-tier multiple-choice diagnostic test (see Section 3.2.1).

Procedures, Scoring, and Data Analysis

Before applying data collection in schools and universities, researchers asked permission to administer the tests to related institutions and were granted ethical research approval from the university. With the help and supervision of teachers, the paper-based test was implemented in the classroom. For item scoring, the correct answer was scored as 1 point, and an incorrect answer was scored as 0 points for all the items. Students get 1 point if they address the task correctly in the first and second tiers.

The WINSTEPS version 4.8.0 software (Linacre, 2021b) and Statistical Package for the Social Sciences (SPSS) version 25 (IBM Corp, 2017) were applied in this study. Rasch analysis and some statistical methods such as descriptive statistics, internal consistency using Cronbach alpha were performed in data analysis. All samples in the data set were investigated and included in the data analysis. WINSTEPS software ran the analysis based on joint maximum likelihood estimation equations; in this formulation, we produced item difficulty scores (IFILE) in log odds unit scale (logits) from student raw scores. Logits are interval data ranging from a specific value from negative infinity to a positive infinity number (Linacre, 1998, 2020). Item difficulty data in logits will be used as a data variable to evaluate reliability, validity, the item difficulty pattern, and DIF using Rasch analysis. Rasch analysis has some advantages in explaining the psychometric properties of data such as (1) generating the difficulty level of an item accurately and precisely, (2) detecting the suitability and interaction of items and persons (item–person maps), (3) identifying outliers (person misfit), and (4) detecting item bias (DIF), which is useful for exploring item difficulties' patterns in this study (Boone et al., 2016; Sumintono & Widhiarso, 2014).

4.2.3 Results

Reliability and Validity

Rasch analysis provided two parameters of reliability: item reliability and person reliability, ranging from 0 to 1. Both the item and person reliability are acceptable in this study at 1.00 and 0.8, respectively (Fisher, 2007), and the item internal consistency using Cronbach's alpha value for all items is 0.88 (Taber, 2018). Item reliability is considered excellent if the value is close to 1 (Fisher, 2007; Sumintono & Widhiarso, 2014). It is possible to achieve if a stable item measure is used for measuring stable person measure above 500, the minimum criteria are 30 items for measuring 30 participants that can generate statistically stable measures with 95 % confidence and ±1.0 logits (Azizan et al., 2020). These results establish that the instrument used is sensitive enough to differentiate students' ability on different levels.

Validation criteria based on item fit statistics, standardized mean square residual (ZSTD), and the mean square residual (MNSQ) indicated two items with positive point biserial correlations (PTMA) values: BIO21 (.17) and CHEM23 (.08) do not meet the fit criteria with an outfit MNSQ above 1.6. The ideal outfit and infit MNSQ are 1 based on the Rasch measurement model, but the acceptable values range from 0.5 to 1.5 (approximately 1.6 still acceptable) and infit and outfit ZSTD ranging from −2 to +2 sequentially (Andrich, 2018; Bond et al., 2020). If the MNSQ parameters are acceptable, then ZSTD can be ignored (Linacre, 2021a). All items have a positive PTMA, which shows that all items contribute to measuring the differences in students' abilities at various levels. Thus, we decided to include all items in the analysis. Figure 10 presents item fit criteria based on infit MNSQ.

For the person who fit criteria, the mean of outfit and infit MNSQ are 0.95 and 1.01, which is close to the ideal threshold around 1, and the mean of infit and outfit ZSTD are -0.1 and 0.1, which are still acceptable. The result from the person fit criteria confirms that participants in this study are fit based on Rasch measurement.

Unidimensionality and Local Independence

The principal component analysis of Rasch (PCAR) was used to evaluate instrument dimensionality. The two-tier multiple-choice diagnostic test was used to assess student misconception in sciece, so we assumed that the unidimensionality criteria as a single factor to measure misconception in science as a latent construct. Based on PCAR, a test only measures a dimension if the minimum variance explained by the measure is > 30 % (Linacre, 1998). Results showed that the variance explained by measures was 38.5%, showing that the developed test met the unidimensionality assumption.

Local independence confirms that the performance of one item is independent of the performance of other items, with the raw residual correlation between pairs of the items < 0.3 (Boone et al., 2013). the items in the test have a residual correlation around 0.1 and 0.28 which means that the assumption of local independence was meet in this study.

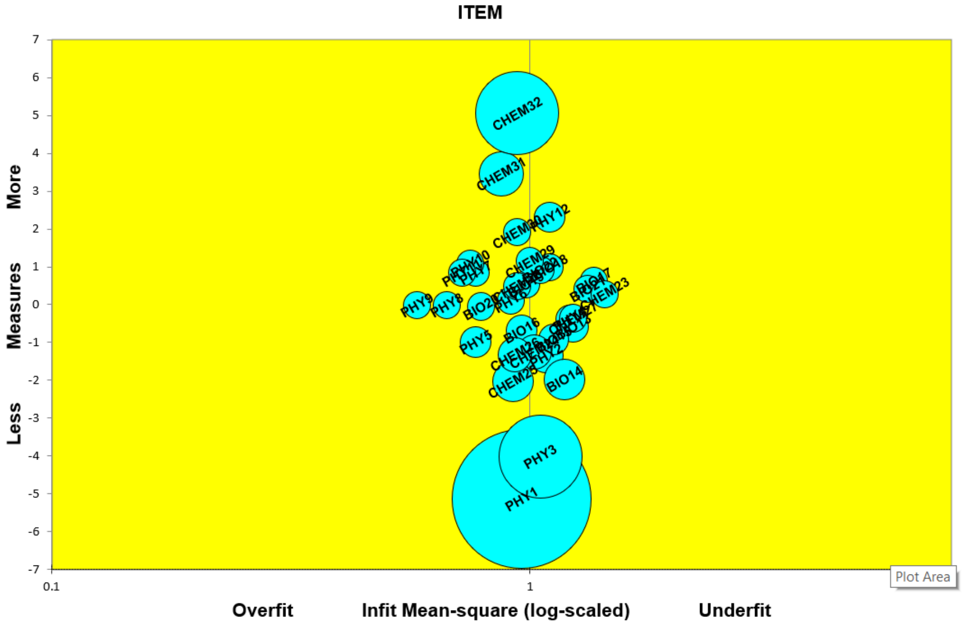


Figure 2. The bubble chart for item fit criteria based on infit MNSQ.

Item difficulty pattern between science concepts and disciplines

We calculated the standard deviation (SD) and the mean of average item difficulty measure for each of the three science disciplines, that is, physics, biology, and chemistry, using item difficulty estimates or logits of items (Table 20). Table 20 shows that the mean of items in chemistry was the most difficult than the mean of items in physics and biology. The mean of items in biology was placed as the easiest on the basis of the mean of item difficulties.

Additionally, we also calculated the item difficulty estimates (measure) on the basis of the 16 science concepts as shown in Table 21 in this study. When comparing item difficulty for each concept, the redox reaction (CHEM 32) with 5.06 logits was the most challenging item to solve among all of the items in chemistry, and kinetic energy (PHY1) with −5.13 logits was the easiest item among all of the items in physics. We explore the specific item difficulty estimates for each item number and item fit parameters in Table 21. Figure 10 also represents the item difficulty pattern in specific science concepts to make it easier to understand data distributions of item difficulty levels between the science concepts and the science disciplines.

Table 17. Standard deviation and mean of item difficulty based on the science discipline (N=32).

Science discipline	Number of items	Difficulty	
		M	SD
Physics	12	−0.56	2.12
Biology	10	−0.07	0.95
Chemistry	10	0.74	2.23

Table 18. Item difficulty estimates and item fit parameters (N=32).

Item code	Discipline	Science Concept	Measure (logits)	INFIT MNSQ	OUTFIT MNSQ	PTMA	Source referenced
PHY1	Physics	Kinetic energy	−5.13	0.96	0.13	0.22	(AAAS, 2019)
PHY2		Kinetic energy	−1.35	1.08	1.06	0.37	Authors
PHY3		Thermodynamics—thermal energy	−4.02	1.05	0.43	0.23	Authors
PHY4		Thermodynamics—thermal energy	−0.38	1.21	1.43	0.28	Authors
PHY5		Impulse and momentums	−0.99	0.77	0.61	0.63	Authors
PHY6		Impulse and momentums	0.11	0.91	0.92	0.52	Authors
PHY7		Atoms and molecules	0.84	0.77	0.71	0.61	(AAAS, 2019)
PHY8		Atoms and molecules	−0.01	0.67	0.59	0.72	Authors
PHY9		Force	−0.02	0.58	0.51	0.78	(AAAS, 2019)
PHY10		Force	1.09	0.75	0.65	0.62	Authors
PHY11		Light	0.85	0.72	0.63	0.66	(Csapó, 1998)
PHY12		Light	2.31	1.10	1.14	0.23	Authors
BIO13	Biology	Cells	−0.59	1.23	1.38	0.27	(AAAS, 2019)
BIO14		Cells	−1.97	1.18	0.66	0.36	Authors
BIO15		Breathing	−0.92	1.12	1.52	0.33	(AAAS, 2019)
BIO16		Breathing	−0.68	0.96	1.27	0.44	Authors
BIO17		Microbes and disease	0.63	1.36	1.34	0.16	(AAAS, 2019)
BIO18		Microbes and disease	0.99	1.10	1.06	0.34	Authors
BIO19		Human body systems	0.53	0.98	1.00	0.45	Authors
BIO20		Human body systems	−0.05	0.79	0.71	0.63	Authors
BIO21		Feeding relationships	0.42	1.32	1.72	0.17	Authors
BIO22		Feeding relationships	0.91	1.05	1.02	0.38	(Csapó, 1998)
CHEM23	Chemistry	Substances and chemical reactions	0.28	1.43	1.68	0.08	(AAAS, 2019)
CHEM24		Substances and chemical reactions	−1.25	1.02	0.92	0.43	Authors
CHEM25		Chemical compound	−2.03	0.92	1.25	0.37	Authors
CHEM26		Chemical compound	−1.32	0.93	0.87	0.48	Authors
CHEM27		Chemical equilibrium	−0.36	1.23	1.47	0.26	Authors
CHEM28		Chemical equilibrium	0.49	0.94	1.00	0.48	Authors
CHEM29		Hydrocarbons	1.15	1.00	0.97	0.41	(AAAS, 2019)
CHEM30		Hydrocarbons	1.92	0.94	0.79	0.41	Authors
CHEM31		Redox reaction	3.46	0.87	0.71	0.31	Authors
CHEM32		Redox reaction	5.06	0.94	0.32	0.20	Authors

A two-way Analysis of Variance (ANOVA) was used to analyze the effect of science concepts and science discipline on item difficulty estimates based on logits. The 2 × 2 ANOVA group in this study achieved the assumption of homogeneity variances based on Levene's test ($p > 0.05$). To validate the normality data assumption, the Kolmogorov–Smirnov test was run before conducting the two-way ANOVA. The results showed that the item difficulty estimates did not differ significantly from normality ($p > 0.05$) with kurtosis (2.21) and skewness (−0.14).

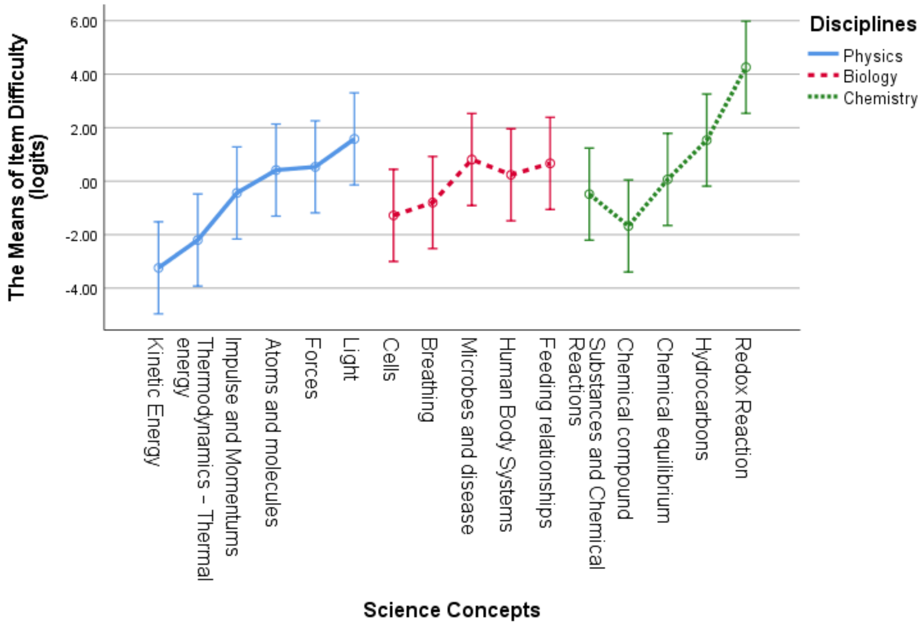


Figure 3. Item difficulty patterns between science concepts and across science disciplines.

As presented in Table 22, the results showed a significant effect of science concepts on item difficulty estimates with a large effect size, F (13) = 4.76, $p < 0.0$. Also, the interaction effect of science disciplines and science concepts showed a significant effect on item difficulty estimates F (15) = 4.59, $p < 0.0$. However, the difference of item difficulties estimates among science disciplines was found to be insignificant, F (2) = 1.30, $p > 0.05$. We can assume that there were no significant differences in the population average among the three different science disciplines, i.e., physics, biology, and chemistry, based on a two-way ANOVA, although the difference in the mean logits of item difficulty as shown in Table 21, positioning items in chemistry as being more difficult than items in physics and biology. Both the science concepts and science disciplines can explain 81% of the variance on item difficulty estimates. To sum up, these findings indicated that the item difficulties pattern varies across science concepts, although there are no significant mean differences of item difficulties among disciplines.

Table 19. Two-way ANOVA for item difficulty measure.

Dependent variable	Sum of squares	*df*	Mean square	F	*p*
Disciplines	9.27	2	4.63	1.30	0.28
Science concepts	81.66	13	6.28	4.76	0.00
Disciplines * Science concepts	90.93	15	6.06	4.59	0.00

R^2 =.81 (adjusted R^2 =.63)

Specific Investigation on Item Difficulty Pattern Among Science Concepts

For understanding concepts in science distributing misconception to students, we can inspect the item difficulty estimates results from Table 5. The item difficulty estimates can be segmented into four categories; very easy (logits < −1), easy (−1 ≤ logits < 0), difficult (0 ≤ logits < 1), and very difficult (logits ≥ 1) (Sumintono & Widhiarso, 2014). Item difficulty estimates in physics showed that concepts of light (PHY11 and PHY12) are more difficult than other concepts in that discipline. All items in physics have logits ranging from −5.12 to 2.13 (very difficult). The concept of kinetic energy (PHY1) is the easiest concept to answer because the concept application can be learned easily. In biology, all item logits are ranging from −1.97 to 0.99. Microbes and disease (BIO 18) have 0.99 of logits (difficult) compared with other items in that discipline, indicating that students have suffered misconceptions and difficulty answering correctly, whereas Cells (BIO 14) is the item that is the easiest one to answer correctly with −1.97 logits. Chemistry has the highest difficulty level among the three science disciplines with logits ranging from −2.03 to 5.06. Redox reaction (CHEM32) has 5.06 logits and was found to be the most difficult item to answer, indicating that students suffer severe misconceptions in redox reaction concepts. To visualize the item difficulty pattern from each concept among disciplines, we calculated the mean of item difficulty pattern for each concept in Figure 12.

DIF based on gender and grade

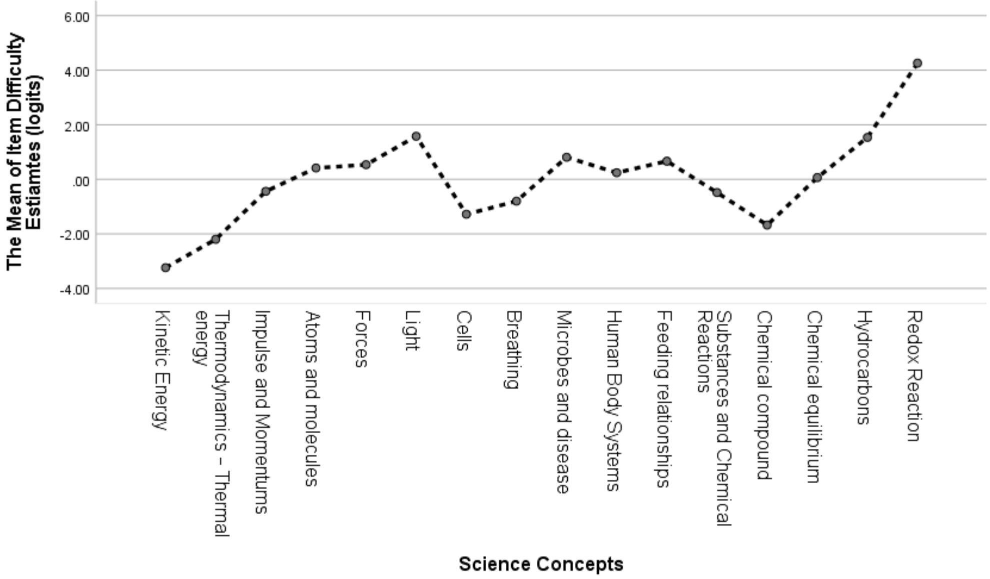


Figure 4. The mean of item difficulty estimates based on science concepts.

DIF analysis was performed to assess differences in item function on the basis of gender and grade on all items in test. DIF analysis investigated item responses on the basis of categorical variables for each item on assessing student misconceptions using a test (Adams et al., 2020; Boone et al., 2013). Differential item functioning analysis is categorized into three types: moderate to large (| DIF | ≥ 0.64 logits), slight to moderate (| DIF | ≥ 0.43 logits), and negligible (Zwick et al., 1999). Figure 13 shows that, overall, items do not have DIF based on gender, except one item in chemistry (CHEM 32). For DIF based on grade, we compared four different cohorts: 10th grade, 11th grade, 12th grade, and the PST. Four items are categorized to differ based on grade: PHY1, PHY5, CHEM23, and CHEM32 (see Figure 14).

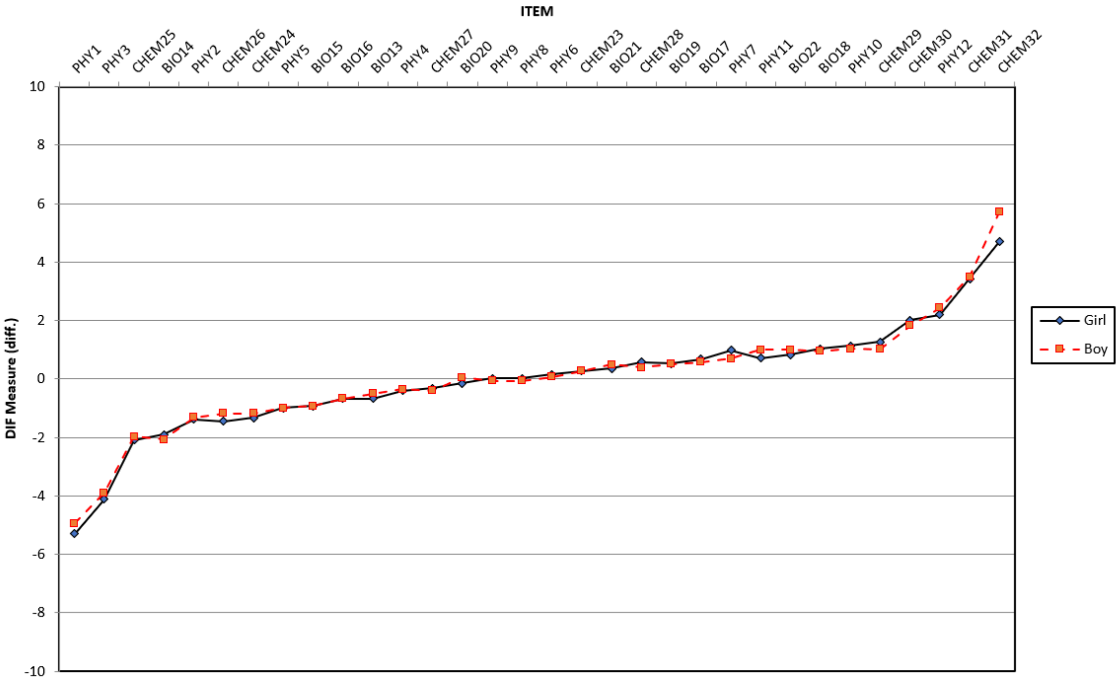


Figure 5. DIF measure based on gender.

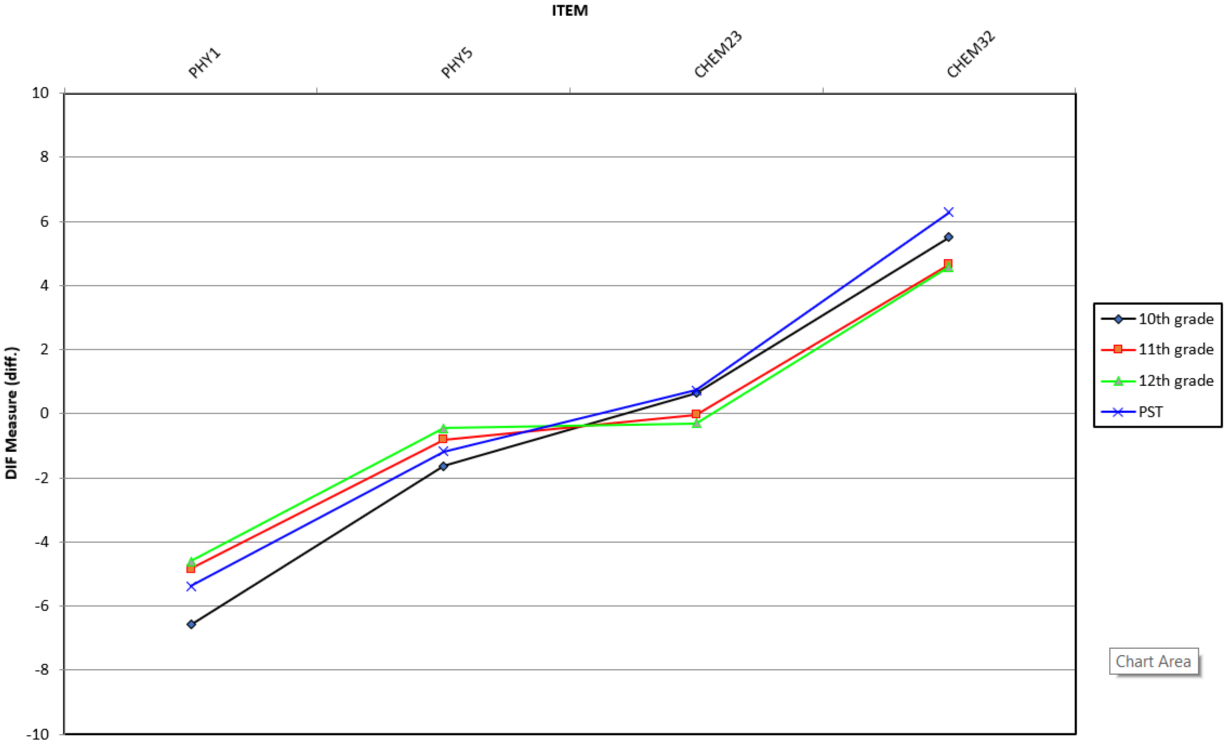


Figure 6. DIF measure based on grade.

4.2.4 Discussion

Through the statistical analysis, we have confirmed that all items used in the developed instrument meet the valid and reliable criteria according to the parameters for the Rasch measurement. The 32 developed items have outfit and infit MNSQ ranging from 0.13 to 1.72 (see Table 22), whereby ZSTD can be ignored if the sample size is more than 500 respondents (Azizan et al., 2020; Linacre, 2021a). Figure 1 shows the item fit pattern based on the MNSQ infit. Several studies had validated the difficulty of items in specific science concepts across science disciplines, such as the concept of energy (e.g., Park and Liu (2019), Neumann et al. (2013)). However, the present study attempts to validate and evaluate item difficulty patterns on various science concepts resulting in student misconceptions that are still limited to the science education area. On the basis of the findings, we can confirm that the item difficulty level is not always reached by students, whereby students must master the more accessible concepts before learning the more complex concepts. This result was in line with previous studies examining the item difficulty level in science subjects (Steedle & Shavelson, 2009), although the science concept under this study is different and the focus is on common concepts causing student misconceptions in science learning.

The difficulty item pattern in the 16 science concepts studied had different average item difficulty levels based on three specific disciplines offered in Indonesian schools (refer to Table 21). The average value of items in the field of chemistry (M: 0.74 logits, SD: 2.23) was much higher than items in the concept of physics (M: −0.56 logits, SD: 2.12) and biology (M: −0.07 logits, SD: 0.95), whereby items with the redox reaction concept (CHEM32) with 5.06 logits in chemistry are the most difficult items to be understood by students, indicating that students often experience misconceptions of the redox reaction concept. These findings were also supported by previous research by Laliyo et al. (2019) measuring the item difficulty level in the redox reaction concept of 1150 Indonesian students having 1.27 logits with the highest logits measure. This study also assumed that the redox reaction was the concept causing students to experience misconceptions. The concept of the redox reaction is an important topic to understand because the redox reaction helps students understand the phenomena that occur in elements in chemical reactions such as losing and gaining electrons or increasing and decreasing oxidation numbers (Treagust et al., 2014).

The results of the two-way ANOVA show that there is a significant effect on the difficulty estimates of whole items on each science concept, $p < 0.05$. There is also a significant interaction between science concepts and disciplines. However, the item difficulty estimates did not differ significantly in the three different science disciplines, $p > 0.05$. These findings are consistent with previous studies that found the item difficulty estimates in science concepts did not differ by science disciplines (Park & Liu, 2019). This finding implies that students' understanding of various science concepts has a different pattern. However, it tends to be similar across science disciplines, especially in physics, biology, and chemistry, indicating that students have different abilities in solving science problems regarding science concepts.

Table 20. The science concept categorization of item difficulty estimates based on the logits mean**.**

Very easy (logits < −1)	Easy item (−1 ≤ logits < 0)	Difficult item (0 ≤ logits < 1)	Very difficult item (logits ≥ 1)
Kinetic energy, thermodynamics—thermal energy, cells, and chemical compound	Impulse and momentums, breathing, microbes and disease, substances, and chemical reactions	Atoms and molecules, feeding relationships, human body systems, and chemical equilibrium	Force, light, hydrocarbons, and redox reaction

To investigate the item difficulty estimates for each science concept in the present study, we categorized the average item difficulty estimates for each concept into four categories in Table 23. Four concepts occupy the very difficult categories, namely, forces, light, hydrocarbons, redox reaction. The forces and light concepts in physics subject were also identified as concepts that distribute misconception to students (Kaltakci-Gurel et al., 2017; Soeharto et al., 2019). In chemistry, the hydrocarbons and Redox were also reported as concepts that were difficult to understand, thus causing student misunderstanding in science learning (Erman, 2017; Laliyo et al., 2019; Ramirez et al., 2020). Five concepts are in the difficult category (see Table 7), specifying students' difficulty in answering or understanding the particular science concept correctly. The item difficulties of each concept were also proven to differ in a previous study by Park & Liu (2019) that reported the item difficulties of the concept of energy concepts in science varied based on students' abilities. Mapping the level of items in science concepts can help teachers realize conditions in teaching specific science concepts considered difficult to learn in classroom activities. By understanding the difficulty level of items in various science concepts, the teacher can estimate which concepts cause students to experience misconceptions in science learning.

DIF confirms that CHEM32 has differences based on gender. In CHEM 32, the item difficulty estimates for females, DIF measure, is 4.69 logits, and for males, the DIF measure is 5.70. These results were in line with previous studies by (Wyse & Mapuranga, 2009) that reported that DIF might happen based on the respondent background, such as gender, and the DIF measure varies according to the item difficulty level. Hence, the DIF contrast is 1.01 logits indicating females are 1.61 logits less able to address item CHEM 32 than males, so CHEM32 was categorized as moderate to large on DIF. DIF based on grade confirmed that four items were difficult for students to understand based on the school level: PHY1, PHY5, CHEM23, and CHEM32. These findings indicate that the school level or grade has a reasonably significant implication in assessing the differences in students’ ability to work on items on science concepts. Comparing the DIF contrast from 10th grade to 11th grade, 12th grade, and the PST for PHY1, PHY5, and CHEM32, the DIF contrast on PHY1 was categorized into moderate to large DIF with 1.73 logits, 1.99 logits, and 1.28 logits, respectively, showing that students in the 10th grade were less able to solve PHY1 than the other grades. The DIF contrast on PHY5 was categorized into moderate to large DIF with 0.83 logits, 1.18 logits, and 0.46 logits showing students in the 10th grade were less able to solve PHY5 than the other grades. The DIF contrast on CHEM32 was categorized into moderate to large DIF with −0.84 logits, −0.93 logits, and 0.77 logits indicating that students in the 10th grade can better solve item CHEM32 than those in the 11th and 12th grades, but those in the 10th grade have less ability than the PST to solve item CHEM32. The DIF contrast on CHEM23 was categorized into moderate to large DIF for 11th–10th grades (−0.676 logits) and 12th–10th grades (−0.943), the negative values showing that students in the 11th and 12th grades have less ability to solve item CHEM23 than those in the 10th grade.

4.2.5 Conclusion

In summary, all items in the developed two-tier multiple choices diagnostic test meet the valid and reliable criteria. Our study confirms that the difficulty level of items on various science concepts is not universally based on science topics, but they are connected or similar across science disciplines, especially in physics, biology, and chemistry. We also found particular items in the science concept may have different difficulty levels based on gender and grade.

4.3 Exploring Indonesian student misconceptions in science concepts

4.3.1 Introduction

Students continuously develop attributes like knowledge, attitudes, and experiences to learn new scientific concepts based on their interactions with the environment and construct their understanding of science by incorporating such attributes into their learning activities. In some cases, the construction of science-related concepts may lead to an incorrect grasp of these ideas, which persists even after learning in science class (Eshach et al., 2018; Prodjosantoso et al., 2019; Stefanidou et al., 2019). Allen (2014) also stated that students experience misconceptions in formal and informal settings unrelated to scientific knowledge. Moreover, students' misconception of science concepts is also triggered by the continuous development of science and technology; consequently, the meanings of science concepts change (Kaltakci-Gurel et al., 2017; Kiray et al., 2015). This condition makes conceptual learning an essential topic in science education to improve student achievement in science subjects. Students' misunderstanding or misconception refers to incorrect generalizations associated with their life experiences, teachers' misinformation, student and teacher misconceptions and reflections of misconceptions in science textbooks (Chazbeck & Ayoubi, 2018; Soeharto et al., 2019). Students' understanding of science concepts may be different based on the scientific context and scientific facts; therefore, this study uses the term 'misconception' to represent students' misunderstanding or alternative conceptions.

In Indonesia, Core competencies and learning indicators embedded in the national curriculum. the Ministry of Education and Culture (MOEC) composed the Indonesian national curricula (Curriculum 2013). Teacher has cumulative task to teach student based on the Core competencies and learning indicators in each discipline. Students study science in general at the junior high school level (7th grade to 9th grade). The specific subject in science like Biology, Physics and Chemistry is only taught at the senior high school level (10th grade to 12th grade) (Faisal & Martin, 2019). However, Student misconception in science rarely seems to be assessed on learning activity and tests in school level, or even on Indonesian national examination. Teacher focused on helping students achieving learning indicators based on national curricula without realizing if students may suffer misconception in particular science concepts. Assessment to identifying misconception or alternative conception in science is a pivotal aspect to improve student understanding related to science concepts. The well-constructed student conceptions in science will lead to students' development and achievement in the science education area. This study attempts to investigate students' misconceptions in science at the senior high school level and university level. Our instrument also attempts to map item difficulty level and compare the development of student misconceptions based on gender and grade school.

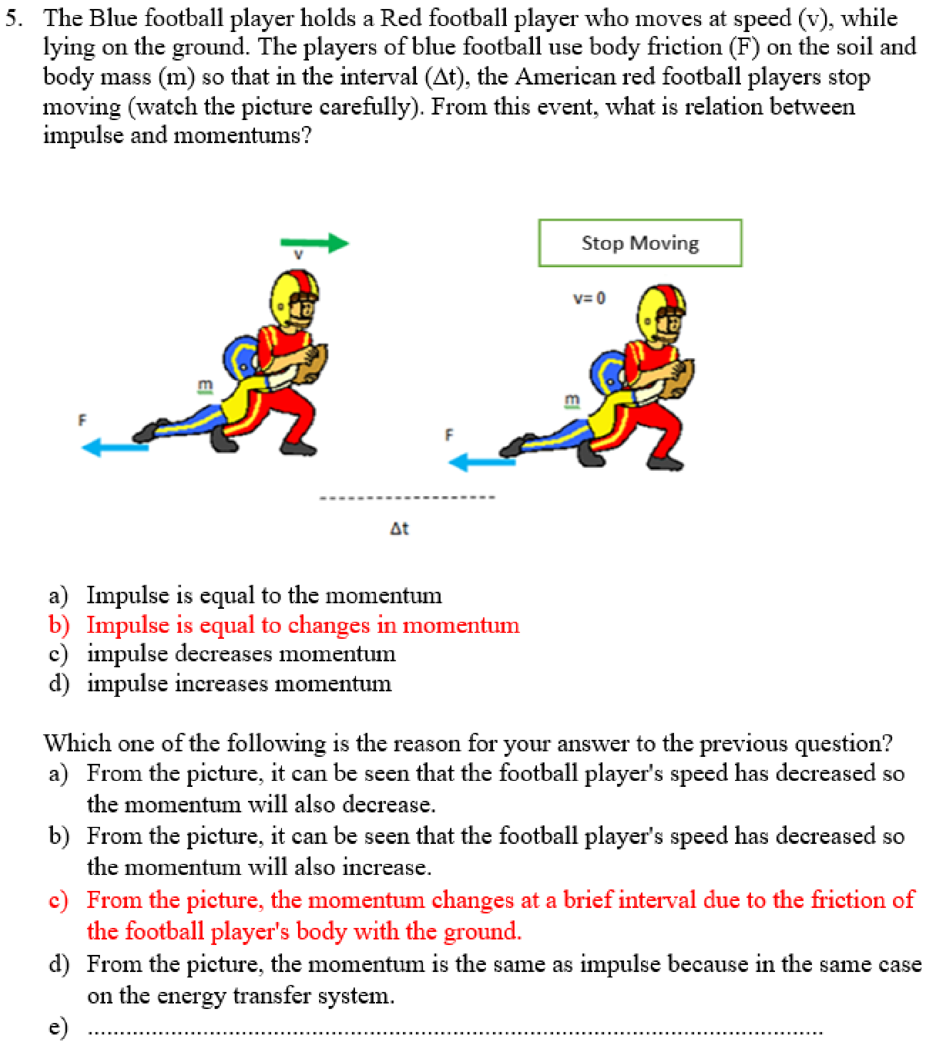
4.3.2 Method

Participants

The participants were recruited via a stratified random sample of 856 students (52.3% females and 47.7% males) from the 10th to the 12th grades at a senior high school and PSTs from three different universities in West Kalimantan Province, Indonesia. The paper-based test was administered, and participants spent 120 minutes completing the test under the supervision of researchers and teachers. However, not all participants were included in the data analysis, as the study applied Rasch modelling for data scaling to filter outliers.

Instruments

The background questionnaire was adapted from the Indonesian version of the PISA 2015 SES instrument (OECD, 2016) (See Section 3.2.1). To capture student misconceptions or alternative conceptions, we implemented the developed two-tier multiple-choice diagnostic test (See Section 3.2.1). All science concepts were adjusted based on the Curriculum 2013 implemented in the Indonesian educational system. Figure 15 represents a sample in the two-tier multiple-choice diagnostic test in Physics.


v
Stop Moving
v= 0
m
F
m
F
Δt

5. The Blue football player holds a Red football player who moves at speed (v), while lying on the ground. The players of blue football use body friction (F) on the soil and body mass (m) so that in the interval (Δt), the American red football players stop moving (watch the picture carefully). From this event, what is relation between impulse and momentums?

a) Impulse is equal to the momentum
b) Impulse is equal to changes in momentum
c) impulse decreases momentum
d) impulse increases momentum

Which one of the following is the reason for your answer to the previous question?
a) From the picture, it can be seen that the football player's speed has decreased so the momentum will also decrease.
b) From the picture, it can be seen that the football player's speed has decreased so the momentum will also increase.
c) From the picture, the momentum changes at a brief interval due to the friction of the football player's body with the ground.
d) From the picture, the momentum is the same as impulse because in the same case on the energy transfer system.
e) ....

Figure 7. A sample item in the two-tier multiple-choice diagnostic test on impulse and momentums in the physics task.

Procedures and data analysis

To collect data, we asked permission to administer the test in the schools and universities, and the Institutional Review Board at the University of Szeged granted ethical research approval. With the guidance and supervision of researchers and teachers, the test was successfully administered. The statistical tools for data analysis included the Statistical Package for the Social Sciences (SPSS) version 25 (IBM Corp, 2017), MPLUS 8.4 (Muthén & Muthén, 2017) and WINSTEPS version 4.7.0 for Rasch measurement (Linacre, 2020). Students' total scores were converted into the log odd unit scale (logits) assumed as interval data ranging from negative to positive infinity. Further, this study performed item–person maps, outlier analysis, model fit analysis, reliability and validity analyses, descriptive statistics, stepwise regressions, t-test, and analysis of variance (ANOVA). All Rasch analysis procedure follows the guideline for Rasch analysis from Linacre (2021a) and Boone et al. (2013).

4.3.3 Results

The findings were derived from the following research analyses: (1) scaling outliers based on misfitting person identification and person diagnostic maps (PKMAPs), (2) finding model fit based on confirmatory factor analysis using unweighted least squares (ULS) estimator and the Rasch measurement for item validity and reliability, (3) Wright maps to present item–person interactions, (4) t-test and ANOVA to measure differences based on gender and grade level and (5) multiple linear regression using the stepwise method to find factors that predict students' science conceptions.

Scaling outliers or misfitting persons

Before performing further analysis, we screened the data for outliers, also known as 'misfitting persons', which refer to student responses that show inconsistency or indicate guesswork. Rasch analysis allows researchers to screen the data for misfitting persons so that the data ascertain the true ability of students' scores to represent their ability to understand scientific concepts. From the dataset, we excluded 102 misfitting students out of 856 which involves 594 students at the senior high school level and 160 students at the university level. data were analysed using Rasch modelling and WINSTEPS version 4.7.0 based on the joint maximum likelihood estimation formula, wherein the raw data were converted into logits as interval data (Linacre, 1998, 2020). Table 24 shows the summary statistics of students and items in this study after excluding misfitting persons.

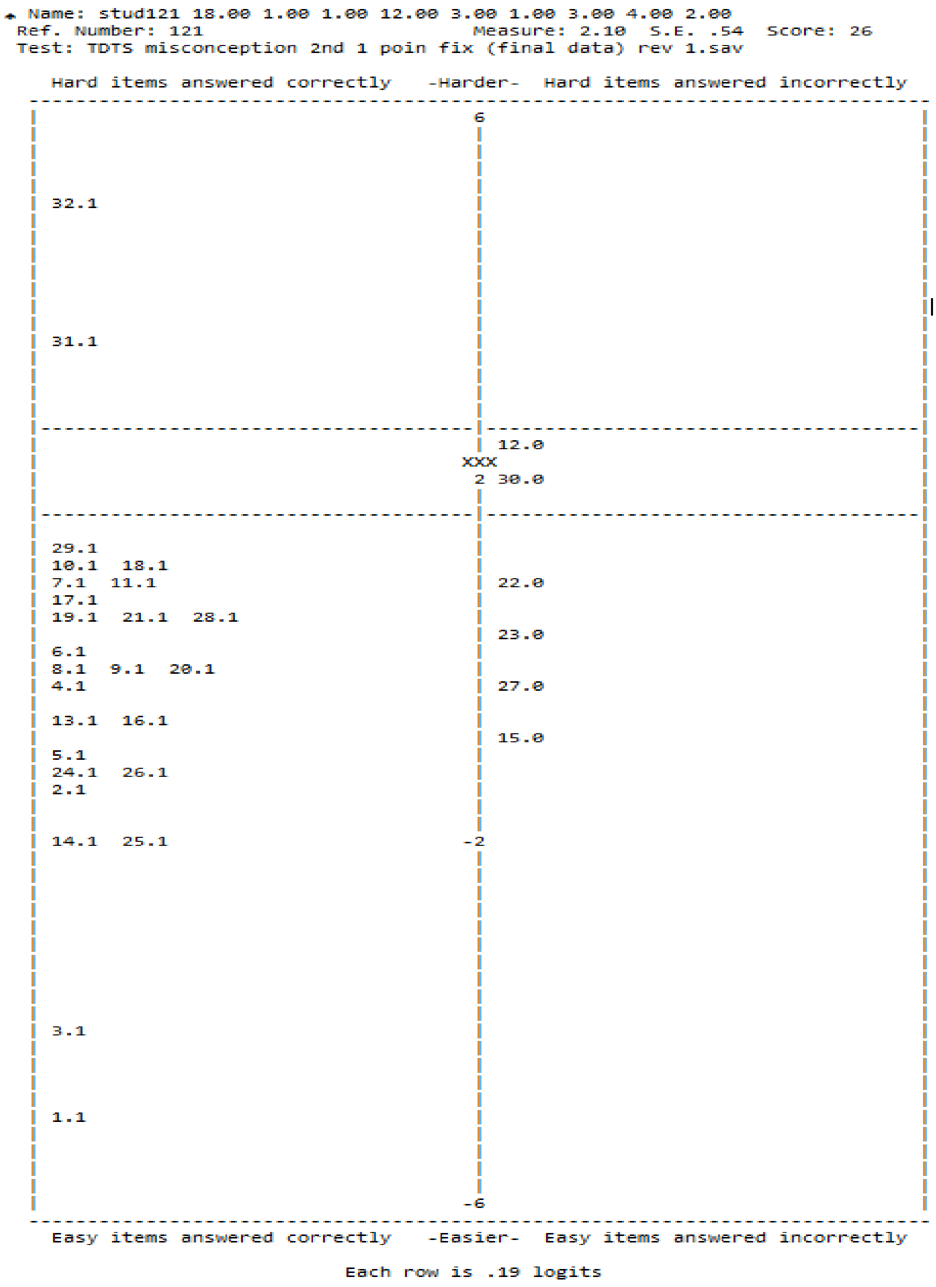
Misfitting students were identified based on person infit and outfit mean of the squared (MNSQ) criteria. If infit and outfit MNSQ values are outside the acceptable range of 0.5–1.5 (around 1.6 still acceptable), the student is included in the misfitting or outlier category (Andrich, 2018; Bond et al., 2020). Another indicator of misfitting students, person infit and outfit z-standardized (ZSTD), has acceptable values ranging from −2 to +2 in sequence (Bond et al., 2020). However, infit and outfit ZSTD can be ignored if the sample size is more than 500 and if the infit and outfit MNSQ criteria have been met (Azizan et al., 2020; Linacre, 2021a).

Table 21. Summary statistics of students and items (N=856).

	Senior high school students		University students	
	Persons	Items	Persons	Items
N	594	32	160	32
Mean measure	0.70	0.00	0.70	0.00
Mean	18.7	454.8	18.7	404.5
SD	0.98	2.32	0.98	2.33
SE	0.49	0.11	0.48	0.12
Mean outfit MNSQ	1	1	1	1
Mean infit MNSQ	0.96	0.96	0.9	0.91
Separation	2	12.34	2	12.45
Reliability	0.80	0.99	0.80	0.92
Cronbach's alpha	0.82			0.88
Raw variance explained by measures	36.1%			30.1%
Chi-squared (χ2)	21716.79 (df = 21746)			212.11 (df = 21746)
Probability	0.5544*			0.64*

*Normally distributed

We adopted PKMAPs to obtain more detailed information on the need for data scaling to detect outliers before further analysis. Stud121, a sample case from the misfitting student category (infit MNSQ: 1.67, outfit MNSQ: 2.19), had inconsistent response patterns in PKMAPs as shown in Figure 2. PKMAPs describe students' ability to respond according to the difficulty level of an item. In Figure 2, the most difficult items are at the top of the diagram, and the easiest ones are at the bottom. Correct student responses are on the left, whereas incorrect ones are on the right. While stud121 correctly answered the two most difficult items, numbers 31 and 32, they were incorrect in the easier items, such as numbers 12, 30 and 22, and such inconsistency in responses might have been due to the student's carelessness. Because the student's correct answers to more difficult items were higher than their logit ability, these responses are considered lucky guesses.



```
◆ Name: stud121 18.00 1.00 1.00 12.00 3.00 1.00 3.00 4.00 2.00
 Ref. Number: 121                       Measure: 2.10  S.E. .54  Score: 26
 Test: TDTS misconception 2nd 1 poin fix (final data) rev 1.sav

    Hard items answered correctly   -Harder-  Hard items answered incorrectly
  -----------------------------------------------------------------------------
  |                                     6                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  | 32.1                                |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  | 31.1                                |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |-------------------------------------|-------------------------------------|
  |                                     | 12.0                                |
  |                                    XXX                                    |
  |                                     2 30.0                                |
  |                                     |                                     |
  |-------------------------------------|-------------------------------------|
  |                                     |                                     |
  | 29.1                                |                                     |
  | 10.1  18.1                          |                                     |
  | 7.1  11.1                           | 22.0                                |
  | 17.1                                |                                     |
  | 19.1  21.1  28.1                    |                                     |
  |                                     | 23.0                                |
  | 6.1                                 |                                     |
  | 8.1  9.1  20.1                      |                                     |
  | 4.1                                 | 27.0                                |
  |                                     |                                     |
  | 13.1  16.1                          |                                     |
  |                                     | 15.0                                |
  | 5.1                                 |                                     |
  | 24.1  26.1                          |                                     |
  | 2.1                                 |                                     |
  |                                     |                                     |
  |                                     |                                     |
  | 14.1  25.1                         -2                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  | 3.1                                 |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  | 1.1                                 |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                     |                                     |
  |                                    -6                                     |
  -----------------------------------------------------------------------------
    Easy items answered correctly   -Easier-  Easy items answered incorrectly

                             Each row is .19 logits
```

Figure 8. Responses by stud121 based on PKMAPs.

Wright map based on grade

The Wright map in Figure 17 illustrates the interaction between student ability and item difficulty based on grade. Item difficulty level is on the right side of the map, whereas student abilities based on four categories (10th grade, 11th grade, 12th grade and PST) are on the left side. The logit value determines the item's difficulty level (Boone et al., 2013): the higher the item logit, the more difficult the correctly answered item, and the lower the item logit, the easier the correctly answered item. Figure 17 shows that the most difficult item to correctly answer is item 32 (redox reaction concepts) in the chemistry task, whereas the easiest item to correctly answer is item 1 (kinetic energy concepts) in the physics task. Simultaneously, the Wright map evaluates student ability and item difficulty level using the same linear interval scale of item measure (logit). In addition, we found that student ability did not show significant differences for each grade level, indicating that the students experienced persistent misconceptions in science. Through the Wright map, we were able to evaluate how items and persons corresponded to the theoretical prediction.

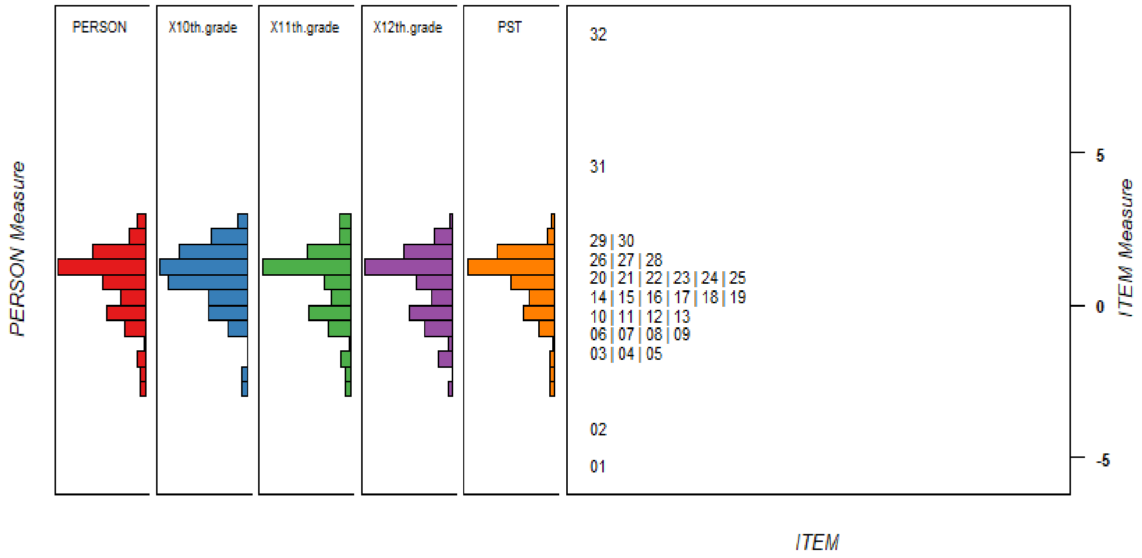


Figure 9. Wright item–person map based on grade levels.

Item reliability

Table 24 demonstrates that the internal consistency was assessed using Cronbach's alpha for all items and the item–person reliability parameter based on Rasch analysis. The Cronbach's alpha value for all items was 0.80, indicating high internal consistency and reliability (Taber, 2018); hence, all items were retained. Meanwhile, the Rasch model showed good person and item reliability values, which were 0.80 and 0.99, respectively (values higher than 0.67 indicate good reliability) (Fisher, 2007). Generally, in terms of reliability indicators, the two-tier multiple-choice diagnostic test met the acceptable threshold.

Confirmatory factor analysis for model fit

One of the best measures for the construct validity of a research instrument is CFA. To perform CFA, we employed MPLUS 8.4 (Muthén & Muthén, 2017) with two

CFA models with the ULS estimator, as it provides more accurate results regarding standard errors, estimates and fit indices than weight least square (WLS) or maximum likelihood (ML) (Muthén, 1993). CFA evaluated the model based on standardized root mean square residual (SRMR), comparative fit index (CFI) and the root mean square error of approximation (RMSEA). Goodness-of-fit indices measured how well the rotated matrix matched the original matrix. CFI required a large number of values and compared the real correlation matrix with the reproduced correlation matrix. RMSEA and SRMR pertain to the value of residual statistics, which are expected to be small in the residual matrix. Hence, we observed the following cut-off values to assess model fit: SRMR < .08, CFI > .90 and RMSEA < .06 (Caleon & Subramaniam, 2010; Hu & Bentler, 1999).

The first model proposed a single-factor CFA model with 32 items in a single group; this model showed acceptable goodness-of-fit indices. The results showed that all cut-off criteria values were met, all of which had a significant positive factor loading [CFI = .973, RMSEA = .006, CI (.001, .014) and SRMR = .017]. The second model proposed a three-factor CFA model based on biology, chemistry, and physics tasks. The results showed that all cut-off criteria values were met and were less than the single-factor model [CFI = 0.939, RMSEA = .010, CI (.01, .017) and SRMR = .017]. Overall, the single-factor model showed the best fit, indicating acceptability in terms of construct validity and achieving unidimensionality in a single factor.

The measurement invariance was conducted to compare based on senior high school and university level through CFA in measurement models to confirm the measurement model in this study measures the same underlying latent construct across the different groups. In other words, the instrument is not different if we measure two group levels, students from senior high school and university level. We found that there were no significant differences between senior high school and university levels in terms of group invariance. The invariance testing showed there are no significant invariances when comparing from Metric against Configural (Chi-square =1.971, p= 0.9223), Scalar again Configural (Chi-square =10.273, p= 0.5920), and Scalar against Metrics (Chi-square =8.302, p= 0.2168).

Rasch analysis for item fit

The criteria applied to validate item-level appropriateness include the infit and outfit MNSQ, infit and outfit ZSTD and point-biserial correlations (PTMA). However, we excluded infit and outfit ZSTD because the sample size was more than 500 (Linacre, 2021a). For infit and outfit MNSQ, the acceptable range is 0.5–1.5, with about 1.6 still acceptable (Andrich, 2018; T. G. Bond et al., 2020; Boone et al., 2013). All test items had positive PTMA, which evaluates whether items function according to the intended model in measuring a construct. PTMA was used as an additional threshold to confirm item fit. A positive PTMA value indicates that all items are acceptable, but a negative PTMA value shows that an item does not function well when compared with other items (T. G. Bond et al., 2020; Boone et al., 2013). Table 25 shows the results of the Rasch analysis using difficulty level (logit), infit and outfit MNSQ and PTMA. The item fit analysis results in Table 25 indicated that all items met the model fit criteria. Moreover, item separation (see Table 24) had a value of 12.34, indicating various levels of item difficulty, and the person separation value was 2, showing that the test could distinguish at least two groups of students: high and low performance. Therefore, we included all items in the analysis because the infit and outfit MNSQ and PTMA criteria were fulfilled.

Figure 17 and Table 25 show that item 32 (CHEM32) is the most difficult item, but its value is still within the acceptable range based on infit and outfit MNSQ. Notably, however, this item seemed too difficult and needed to be revised to match sample targets; meanwhile, this result also indicated that students at every level have severe misconceptions (0.13% correct answers) regarding redox reactions in chemistry. An item would be considered a misfit only if the three abovementioned criteria (infit MNSQ, outfit MNSQ and PTMA) are not achieved. Generally, we can assume that the collected data used all items in the two-tier multiple-choice diagnostic test from 10th, 11th and 12th graders and PSTs to assess scientific misconceptions matching the Rasch model.

Based on the principal component analysis of Rasch (PCAR), the test has achieved the unidimensionality assumption with the variance explained by measures was 38.5%. The unidimensional test can be achieved if the minimum variance explained by the measure is > 30 % (Linacre, 1998). Items in the test have a residual correlation of around 0.1 and 0.28 confirming item dependency achieved whereby the raw residual correlation between pairs of the items < 0.3 (Boone et al., 2013). The unidimensionality assumption is used to confirm the items in the instrument measure the same construct namely student misconception in science. This procedure follows the Rasch analysis for the unidimensional model using WINSTEPS Software (Boone et al., 2014; Linacre, 2021a).

DIF analysis can be used in several background variables using categorical data in comparing items in a test (Boone et al., 2013). Differential item functioning analysis is categorized into three types: moderate to large (| DIF | ≥ 0.64 logits), slight to moderate (| DIF | ≥ 0.43 logits), and negligible (Zwick et al., 1999). To confirm item bias, the differential item functioning (DIF) analysis was utilized based on gender. The results confirm that all items do not have DIF based on gender. We found one item in chemistry (CHEM 32) with significant probability ($p < 0.01$), but the DIF size can be categorized as negligible, DIF contrast < 0.43).

Table 25. Item fit analysis (N=32).

Item	Science concept	Correct answer (%)	Measure (logit)	Infit MNSQ	Outfit MNSQ	PTMA	
PHY1	Kinetic energy	99.47	−5.34	0.96	0.12	0.20	(AAAS, 2012)
PHY2	Kinetic energy	83.69	−1.29	1.07	1.16	0.35	Authors
PHY3	Thermodynamics–Thermal energy	98.41	−4.18	1.03	0.36	0.23	Authors
PHY4	Thermodynamics–Thermal energy	70.69	−0.34	1.21	1.35	0.28	Authors
PHY5	Impulse and momentum	79.18	−0.91	0.76	0.60	0.64	Authors
PHY6	Impulse and momentum	59.81	0.27	0.93	0.97	0.49	Authors
PHY7	Atoms and molecules	43.24	1.1	0.81	0.75	0.56	(AAAS, 2012)
PHY8	Atoms and molecules	61.27	0.2	0.66	0.59	0.72	Authors
PHY9	Forces	61.94	0.16	0.59	0.53	0.77	(AAAS, 2012)
PHY10	Forces	37.53	1.38	0.80	0.68	0.56	Authors
PHY11	Light	43.10	1.1	0.76	0.67	0.61	Authors
PHY12	Light	20.56	2.35	1.04	0.92	0.27	Authors
BIO13	Cells	72.02	−0.42	1.23	1.45	0.25	(AAAS, 2012)
BIO14	Cells	87.93	−1.73	1.19	0.70	0.36	Authors
BIO15	Breathing	78.51	−0.86	1.05	1.15	0.41	Authors
BIO16	Breathing	73.34	−0.5	0.97	1.29	0.43	Authors
BIO17	Microbes and disease	51.86	0.67	1.36	1.32	0.15	Authors
BIO18	Microbes and disease	39.12	1.3	1.17	1.15	0.25	Authors
BIO19	Human body systems	50.80	0.73	0.98	0.97	0.44	(AAAS, 2012)
BIO20	Human body systems	61.41	0.19	0.82	0.76	0.60	Authors
BIO21	Feeding relationships	50.93	0.72	1.38	2.00	0.06	Authors
BIO22	Feeding relationships	43.10	1.1	1.03	0.98	0.38	Authors
CHEM23	Substances and chemical reactions	57.96	0.37	1.40	1.60	0.10	(AAAS, 2012)
CHEM24	Substances and chemical reactions	80.90	−1.05	1.03	0.95	0.43	Authors
CHEM25	Chemical compound	88.73	−1.82	0.93	1.35	0.37	Authors
CHEM26	Chemical compound	82.10	−1.15	0.96	0.96	0.46	Authors
CHEM27	Chemical equilibrium	70.56	−0.33	1.16	1.25	0.32	Authors
CHEM28	Chemical equilibrium	55.04	0.52	0.86	0.92	0.54	Authors
CHEM29	Hydrocarbons	38.86	1.31	0.95	0.89	0.43	(AAAS, 2012)
CHEM30	Hydrocarbons	24.01	2.12	0.92	0.76	0.39	Authors
CHEM31	Redox reaction	3.85	4.33	0.90	0.57	0.25	Authors
CHEM32	Redox reaction	0.13	8.97	1.00	1.00	0.00	Authors

Differences in students' science misconceptions according to grade level

We performed ANOVA to compare students' conception scores across school grades and PSTs on the test and subtest. No significant differences were observed between students' understanding of science concepts in physics [F (3,750) = 1.83, p > .05] and chemistry [F (3,750) = 1.51, p > .05]. However, we found mean significant differences in the biology subtest [F (3,750) = 3.34, p < .05]. For the whole test, the results showed that student conception mean scores differed between grades [F (3,750) = 2.653, p < .05]. Because equal variances are not assumed based on Levene statistics (p < .05), we performed a Dunnett T3 test for post-hoc analysis to identify differences between cohorts, presented in Table 26.

Table 26 shows that students' conception scores are different between grade levels. Although ANOVA results for the entire test showed significant differences between cohorts, post-hoc analysis showed no significant differences with less than a 5% probability except for the biology subtest for 10th and 11th graders (p = 0.25) and for 10th and 12th graders, which showed substantial differences. This might indicate that student misconceptions are resistant to change, persistent and rooted deeply in science concepts, making it more difficult for higher-level students to understand science. Figure 18 shows that students at higher levels (PSTs) develop higher misconceptions than other cohorts; for instance, Student 272 from the PST cohort correctly answered five of 32 items (around 15%), proving that higher-level students experience higher misconceptions than others.

Table 22. Dunnett T3 multiple comparisons of student conceptions between senior high school students and prospective science teachers (N=856).

Grade	Physics		Biology		Chemistry		Test	
	Mean differences	p	Mean differences	p	Mean differences	p	Mean differences	p
10th & 11th	.52	.24	.51	.02	.19	.83	1.23	.07
10th & 12th	.62	.22	.58	.04	.19	.91	1.40	.09
10th & PST	.26	.93	.35	.37	.10	.99	.72	.69
11th & 12th	.09	.98	.07	.99	−.01	.99	.17	.99
11th & PST	−.25	.94	−.16	.96	−.09	.96	−.51	.92
12th & PST	−.35	.86	−.23	.90	−.09	.98	−.68	.56

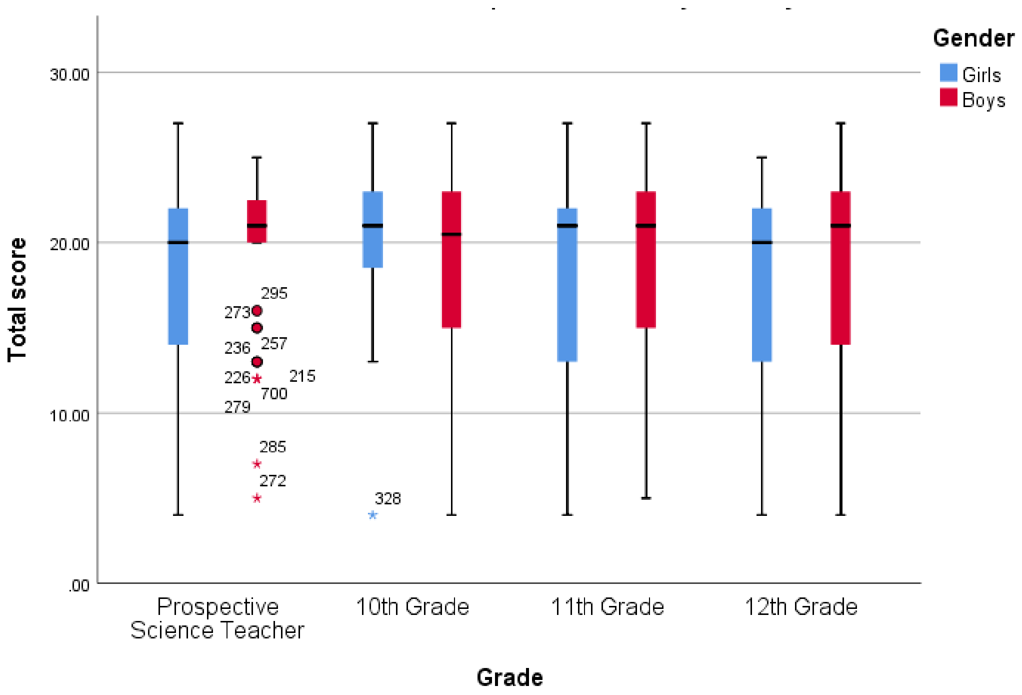


Figure 10. Comparison of student misconceptions between school grades.

Differences in students' science misconceptions based on gender

We conducted an independent-sample t-test to compare students' conceptions in the tests and subtests according to gender. The results showed significant differences in tests and subtests between boys and girls, with mean scores ranging from 4.87 to 19.21 as shown in Table 27. Boys' mean scores for the whole test and subtests were higher than those of girls, showing that boys comprehend science concepts and solve science problems better than girls. In addition, the mean score comparisons showed that the chemistry subtest was more difficult than the other subjects, as the mean scores of boys and girls in that subtest were lower than in the other subtests, confirming the item difficulty (logit) results in Table 25.

Table 23. Independent-sample t-test comparing student conceptions according to gender.

Subject	Girl	Boy		
	Mean (SD)	Mean (SD)	t	p
Physics	7.38 (2.87)	7.80 (2.73)	−2.03	.042
Biology	5.94 (2.07)	6.24 (1.90)	−2.07	.039
Chemistry	4.87 (1.89)	5.16 (1.81)	−2.14	.032
All subjects (science)	18.20 (5.59)	19.21 (5.09)	−2.58	.010

Predicting students' science misconceptions

To evaluate how some factors affect students' misconceptions in science, we performed multiple regression using the stepwise method, in which the predictors are school category, grade level, gender, father's education, mother's education and school performance. The results showed that only the gender predictor could significantly explain 9% of the variance on student misconception mean scores [F (753) = 6.6, $p < .05$]. This indicated that gender was a pivotal factor in predicting the science misconceptions of 10th, 11th and 12th graders and PSTs in Indonesia.

4.3.4 Discussion

The results showed that the two-tier multiple-choice diagnostic test could reliably assess students' misconceptions at the senior high school (10th, 11th, and 12th grades) and PST levels. The test met the criteria for Cronbach's alpha (0.82) (Taber, 2018) and person and item reliability (0.80 and 0.99, respectively) (Fisher, 2007), which meant the test can be used in the same cohort range. The combination of reliability analysis based on internal consistency and item reliability based on Rasch parameters can provide more convincing results for researchers. The two-tier multiple-choice diagnostic test also showed good validity based on unidimensionality criteria, with 36.1% of variance explained by its measures, indicating that the test can evaluate a single dimension of science misconception. Meanwhile, the CFA analysis revealed that the single-factor or one-dimension model had higher fit indices that the three-factor model [CFI = .973, RMSEA = .006, CI (.001, .014) and SRMR = .017]. Infit and outfit MNSQ and PTMA values for all items indicated good item fit. However, the item CHEM32 (redox reaction) seemed too difficult to correctly answer (0.13%) and had a high difficulty level (8.97 logits). When we assessed item fit, we realised that each science concept had different difficulty levels. These findings were consistent with those of (Park & Liu, 2019), who examined item difficulty in several energy concepts. Therefore, we can assume that the two-tier multiple-choice diagnostic test in this study is valid and reliable in evaluating science concepts.

This study employed data screening analysis to identify outliers or misfitting persons. It excluded 102 of 854 students from the dataset because their person misfit parameters were outside the acceptable range (infit and outfit MNSQ, 0.5–1.6). This finding was similar to demonstrations of outlier detection by (2021) in assessing students' thinking ability. This means we can find misfitting persons in each test evaluation including the science context or beyond students' thinking ability. However, researchers have rarely applied outlier detection, especially in science education (e.g. Kaltakci-Gurel et al. (Kaltakci-Gurel et al., (2017), Arslan et al. (2012), Caleon & Subramaniam (2010), Kiray & Simsek (2021), Peşman & Eryılmaz (2010))

Meanwhile, the Wright map depicted the construction and interaction between student conceptions in science and all items in terms of difficulty level (logit). The constructions of the two-tier multiple-choice diagnostic test covered all student ability levels. However, some items needed revising because they were either too difficult or too easy (e.g. CHEM32, PHY01 and PHY02). Item–person maps showed that students with better ability are more likely to correctly answer more difficult items, whereas those with lower ability are more likely to incorrectly answer such items. When assessing students' misconceptions, science education research usually involves revising items (e.g. (Chan et al., 2021; Laliyo et al., 2019, 2020; Park & Liu, 2019).

Table 26 presents a wide range of students' misconceptions in terms of science concepts and item difficulty level. The concepts in chemistry subjects appeared more difficult than those in other subjects, with correct answers ranging from 0.13% to 88.7%, especially in redox reaction. These results were consistent with those of (2019), Treagust et al. (2014; 2014). Also, ANOVA showed significant differences in student mean scores between cohorts. However, in the post-hoc analysis, we found that significant differences were present only in biology among 10th, 11th, and 12th graders; other subjects in various grade level combinations had no differences in mean scores. Interestingly, PSTs, which are considered higher-level students, experienced higher misconceptions than the other cohorts (see Table 27 and Figure 18). These findings supported the resistant, persistent and deeply rooted nature of science misconceptions (Arslan et al., 2012; Wandersee et al., 1994). Therefore, this study confirmed that higher-level students are more prone to science misconceptions than lower-level ones because of the nature of such misconceptions.

For gender, the independent-sample t-test results confirmed significant differences between girls' and boys' mean scores for subtests and the entire test, which ranged from 4.87 to 19.21, indicating that boys have a higher ability than girls in answering science problems in the test. This is supported by reports that boys are more affected by science motivation and parental role in achievement tests (Taskinen et al., 2015). A study by (Prodjosantoso et al. (2019) also found that boys have higher ability than girls in understanding science concepts. While the stepwise multiple regression results showed gender as the pivotal factor in predicting students' misconceptions in this study, this does not dismiss the possible effect of other factors on students’ science misconceptions, such as textbooks, teacher knowledge and students’ mathematical abilities as described in a review of common science misconceptions by Soeharto, et al. (2019).

The findings of this study have some implications for science teachers in the class context. Students have carried science misconceptions in grade and gender in particular science concepts. Teachers can use the findings to prepare the lesson plans for specific science concepts that will distribute misconceptions often to tackle student learning difficulties. Researchers can explore further how certain science concepts distribute misconceptions to students specifically with various research content. We also hope this study can lead other researchers to utilize Rasch measurement to identify student misconceptions in science.

4.3.5 Conclusion

To conclude, all items in the two-tier multiple-choice diagnostic test met reliability and validity criteria based on CFA and Rasch parameters. Rasch analysis helped to detect misfitting persons or outliers, that is, students with inconsistent answers and lucky guesses. We expect this new method to be used by other researchers before performing further data analysis. Meanwhile, the Wright map showed the interaction between persons and items. However, because the item CHEM32 was considered too difficult and unsuitable for these cohorts, it must be revised if further tests are to be conducted. Further, we confirmed significant differences in student conception mean scores between all cohorts; however, post-hoc analysis results evinced that differences were present only among 10th and 11th graders, and 10th and 12th graders in the biology subtest. In addition, the independent-sample t-test results confirmed that boys’ and girls’ mean scores were significantly different in that the former had higher mean scores than the latter, which demonstrated that boys tend to demonstrate better comprehension of science concepts and can solve science problems better than girls. Multiple linear regression results also identified gender as an essential factor in predicting students’ science misconceptions.

5.4 Assessing Indonesian students' inductive reasoning: Rasch analysis

4.4.1 Introduction

Inductive reasoning is an ability that children need to promote their cognitive development and develop their intelligence. Inductive reasoning facilitates individuals to understand the abstractions of basic rules through their application to various fields of work. Inductive reasoning theories have been integrated from various disciplines, including mathematics, science, philosophy, psychology, and artificial intelligence (Perret, 2015; Sosa-Moguel & Aparicio-Landa, 2021). Worldwide 21st-century competency frameworks have also regarded inductive reasoning components as imperative. Inductive reasoning components are embedded in creativity and problem-solving in various significant competencies such as those found in computer science and technology, communication, social skills, collaboration, and teamwork. Employers, researchers, and policymakers have exhibited an interest in enhancing these related competencies (Chu et al., 2017; Van Vo & Csapó, 2020; Zhu & Neupert, 2021). Reasoning skills play an essential role in workplace environments and various educational fields. The ability to reason inductively is needed and has been learned and is relevant in recent social changes. Furthermore, it is a factor that predicts student outcomes and achievements.

Many studies have revealed inductive reasoning is of importance in various contexts. A relationship between students' inductive reasoning and problem-solving ability and academic success has been found (Csapó, 1997; Csapó & Molnár, 2019; Korom et al., 2017; Perret, 2015; Sosa-Moguel & Aparicio-Landa, 2021). Wu and Molnár (2018) found that inductive reasoning can predict students' interactive problem-solving ability, which is a multi-dimensional cognitive process in specific thinking skills. Furthermore, inductive reasoning can assist student decision-making and be employed to establish various causal relationships (Lafraire et al., 2020; Leighton & Sternberg, 2003). It appears that in 5th, 6th, 7th, 9th, and 11th grade students in the Vietnam context, inductive reasoning continues to increase (Van Vo & Csapó, 2020). However, this study did not reveal how inductive reasoning develops in students at higher levels, for example, by comparing senior high school students with those at undergraduate levels. Measuring inductive reasoning among students at higher levels is crucial to address the lack of information related to the development of students' inductive reasoning. The evaluation of inductive reasoning may be beneficial for evaluating curriculum success, supporting student cognitive development, and providing an overview of student success rates for educators, especially for those who are about to complete their education and start working. The objective measurement such as Rasch analysis can be applied to explore further information related to student and item investigation to assess inductive reasoning. Based on the literature review, we cannot find the application of Rasch analysis for assessing inductive reasoning in the indonesian context whereas Rasch analysis can help researchers to extend the result and literature because Rasch analysis has several advantages such as Rasch analysis completes the requirements of fundamental measurement to transform raw data to a linear interval scale (logits), Rasch analysis allows researchers to investigate student performance and item difficulty using item-person maps, and Rasch analysis is a psychometric technique developed to improve measurement accuracy whereby researchers can construct instruments and monitor instrument quality (Boone, 2016; Kleppang et al., 2020; Tavakol & Dennick, 2013).

In Indonesia, thinking skills were included in the 2013 Indonesian core curriculum (Hasan, 2013; Prastowo & Fitriyaningsih, 2020). This curriculum supported three main domains: attitude, skills, and knowledge in which the learning material was designed to relate to core competencies in different disciplines (Hasan, 2013). This curriculum has a crucial problem regarding assessment methods especially in evaluating attitude assessment. The attitude assessment was completely new and difficult to adapt to the classroom context. Badaruddin & Hawi (2022) reported that the majority of teachers complained about being difficult to assess student attitude, and the teachers' knowledge in choosing the method and developing the assessment instrument was still lacking. However, assessing knowledge and skills was easy (Natsir et al., 2018). To enhance students' thinking skills, the teacher may employ various learning models on different materials and subjects (Prastowo & Fitriyaningsih, 2020). Although inductive reasoning is not taught and trained directly in schools, the inductive reasoning test has been utilized in the general basic skills knowledge test when applying for jobs at the government and company levels. Thus, limited data have been employed to describe students' inductive reasoning at schools and even higher education institutions. Additionally, some limitations in data collection may occur related to internet connection and computer laboratory. Consequently, we use online and paper-based tests, and this method is also used to confirm bias issues regarding the data collection method. Therefore, it is imperative that studies related to the evaluation of inductive reasoning in the Indonesian context are conducted. It is possible that because inductive reasoning is closely related to mathematics, reading, and science performance (De Koning et al., 2002; Nikolov & Csapó, 2018), students' low inductive reasoning ability resulted in Indonesia's low ranking in the 2018 PISA report in which the country was placed 71st out of 77 participating countries (OECD, 2020). Because inductive reasoning was not embedded in the Indonesian core curriculum directly, only a paucity of studies related to inductive reasoning in school and educational contexts have been conducted during the past ten years (Istikomah et al., 2017; Siswono et al., 2020). No studies have employed objective measurements such as Rasch analysis to evaluate students' inductive reasoning in Indonesian. Accordingly, Rasch analysis was utilized in this study to evaluate students' inductive reasoning skills so as to validate an adapted inductive reasoning test in the Indonesian context and classify the difficulty of inductive reasoning items and students' inductive reasoning abilities. It is expected that the current study will form the foundation of research to explore the level of Indonesian students' inductive reasoning and provide information to support teacher and student development.

4.4.2 Method

Participants

A cross-sectional research design with a quantitative method was employed in this study. Stratified random sampling was utilized to select 856 students in the 10th to 12th grade in senior high schools and undergraduate students at universities in West Kalimantan province, Indonesia. Students provided written consent before completing the inductive reasoning test. The anonymity of the students was assured to protect their personal identification. The participants were given 50 minutes to complete the inductive reasoning test under teacher surveillance and guidance during regular class time. The demographic profile of the participants is presented in Table 28.

Table 24. The demographic profile of participants in this study (N=856).

Grade	N	Female/Male ratio (%)	Mean age (years)
10th	231	29.4/70.6	16.02
11th	291	67.7/32.3	17.11
12th	153	41.2/58.8	17.99
Undergraduate students	181	66.3/33.7	19.17
Total	856		

Instruments

Inductive reasoning test

The inductive reasoning test (see Section 3.2.3) was adapted and employed in this study from original version (Csapó, 1997; Pásztor, 2016).

Procedures

The data collection at both the schools and universities was conducted before the COVID19 pandemic in Indonesia by using online and paper-based tests. Csapó et al. (2009) confirmed no bias effect or significant differences between online- and paper-based tests. However, to ensure this effect, we also employed DIF based on online-based and paper-based tests. The researchers obtained ethical approval from the Institutional Review Board at the University of Szeged. Permission was also sought from the school and university to administer the test. The Electronic Diagnostic Assessment System (eDia) platform, which was developed by the Center for Research on Learning and Instruction at the University of Szeged, Hungary (Csapó & Molnár, 2019), was utilized to disseminate the test. The eDia platform can be used to assess various types of tests such as multiple-choice tests, open-ended questions, drag-and-drop items, and feedback answers. The eDia is easy to use as a diagnostic instrument platform that compiles item banks to support teaching and learning in a digital pedagogy system. The participants were able to access the eDia platform through Google Chrome, Mozilla Firefox, and other standard internet browsers. Where the schools and universities lacked the necessary infrastructure and technological support, the paper test was employed. We collaborated with the teachers in observing and giving guidance when finalizing the inductive reasoning test.

Data analysis

The SPSS version 25 (IBM Corp, 2017) was employed to perform descriptive statistics, Cronbach's alpha (α), and McDonald's omega (ω). Furthermore, R software version 1.4.1717 (R Core Team, 2020) with graphical packages such as the yarrr (Phillips, 2017) and the ggplot2 (Wickham, 2016) were used to depict the interactive pirate plot of the participants' inductive reasoning development. WINSTEPS version 5.1.4 software (Linacre, 2021b) for Rasch measurement was utilized to perform data analysis. Rasch analysis included conducting Rasch modelling using joint maximum likelihood estimation (JMLE) in which student scores were converted into the logit scale (interval data), ranging from negative infinity to positive infinity. Rasch parameter evaluation was employed to assess the validity and reliability based on unidimensionality, local independence, and by checking person and item reliability criteria. The Wright map was presented to confirm targeting criteria between item and person. DIF analysis was used to evaluate item bias in accordance with the test method. The logit value of person (LVP) and logit value of item (LVI) were classified using the COUNTIF function in Microsoft Excel in accordance with mean logits and the standard deviation as a recommendation from Chan et al. (2020) and Adams et al. (2020). The COUNTIF function was used to perform automatic calculations of the person ability measure and item difficulty measure based on the mean logits and the standard deviation categorization. This process was done to achieve accuracy in the grouping of persons and items with a large number of respondents and to reduce human errors when grouping manually.

4.4.3 Results

Validity and reliability of inductive reasoning test

Validity

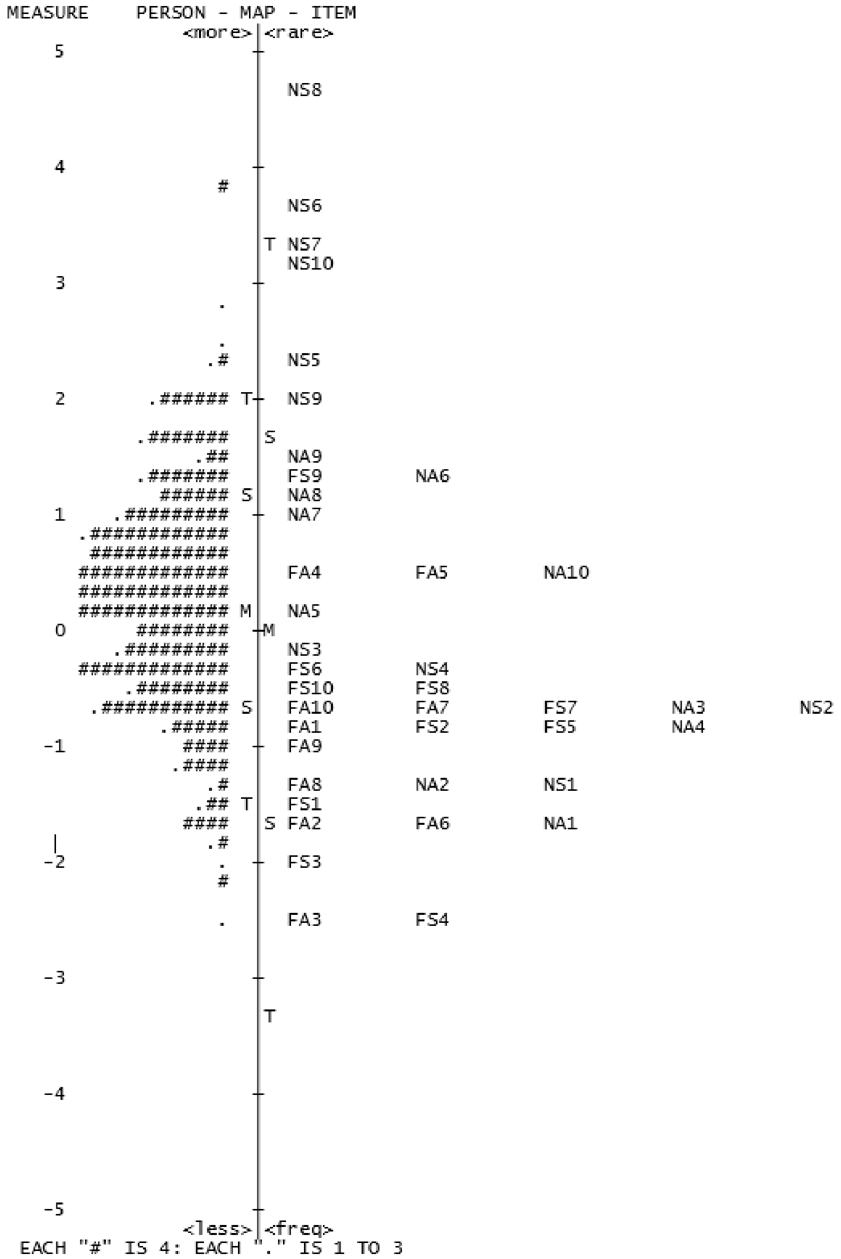
Rasch analysis was performed by employing JMLE estimation for dichotomous data to validate the inductive reasoning test that had been adapted for Indonesia. The item and person parameters were used to validate the inductive reasoning test. Person and item fit validity were identified in accordance with the mean of infit and outfit mean square (MNSQ), where an acceptable range is from 0.5 to 1.5 even though 1.6 is still regarded as acceptable. Furthermore, the ideal values for fit criteria are close to 1.00 logits (Andrich, 2018; Boone et al., 2014). The infit and outfit z-standardized (ZSTD) of persons and items were ignored because the sample was larger than 500 students (Azizan et al., 2020) could differentiate person abilities as latent traits. In addition, item separation revealed that the inductive reasoning test includes a range of easy and difficult items (Boone et al., 2014). It is imperative that separation values should be more than 2 logits in which the larger the separation index, the more superior the quality of the test (Boone et al., 2014; Fisher, 2007; Planinic et al., 2019). The results of Rasch analysis are presented in Table 29. The results confirmed that inductive reasoning for the reasoning test adapted for Indonesia achieved validity in accordance with the Rasch parameter for each task and entire test. It was considered that the FA task met the person separation threshold, with person separation close to 2 logits.

Table 29. The summary of Rasch parameters for inductive reasoning test and task (N=40).

Psychometrics Attribute	Task				IR test
	FA	FS	NA	NS	
Number of Items	10	10	10	10	40
Mean					
item outfit MNSQ	0.95	0.98	1.16	1.54	1.01
item Infit MNSQ	1.00	0.98	0.98	.99	1.00
person outfit MNSQ	0.95	.98	1.16	1.13	1.01
person Infit MNSQ	1.00	.99	0.98	0.96	1.00
Item separation	10.27	12.07	13.62	14.79	16.46
Person separation	1.98	2.18	2.25	2.82	2.92
Unidimensionality					
Raw variance by measure	30.2%	36.6%	36.1%	53.7%	
Unexplained variance 1st contrast	1.72	1.97	1.70	2.03	

WINSTEPS software estimates unidimensional Rasch model, but it also can give benefits to multidimensional model by assessing the subtest (Linacre, 2021a). In this study, we evaluated the task as a subtest as unidimensional model based on the recommendation from Bond and Fox (2015) where the inductive reasoning test was developed to assess an underlying construct that is composed of distinct but related sub-dimensions. Aryadoust and Raquel (2019) also recommended assessing the unidimensionality of the subtest using WINSTEPS when using a test with multidimensionality model as a basic assumption. The unidimensionality and local independence were assessed to confirm the construct validity of the inductive reasoning test. The values of raw variance by measure for all tasks are presented in Table 2. The results revealed that the reasoning test achieved an acceptable threshold of more than 30%. While the unexplained variance for the first contrasting values was less than 2 for all the tasks that confirmed unidimensionality which indicates the test comprised close to four dimensions based on the tasks. Local independence proves that each item in the inductive reasoning test was independent. The raw residual correlation between pairs was also assessed to decide local independence. The acceptable threshold of the raw residual correlation between pairs of items should be less than 0.3 (Boone et al., 2014). The results revealed that all the items in the inductive reasoning test had a residual correlation ranging from 0.11 to 0.28, which supported the assumption of acceptable criteria for local independence.

The interaction between items and students is represented in the Wright map (Figure 19). The Wright map reveals that the items and students matched the targeting criteria. In other words, all the items covered all students' abilities. The results further demonstrated that all the items in the inductive reasoning test met fit criteria based on the infit MNSQ values, ranging from 0.80 to 1.23 logits. While item NS8 (3.34 logits) was indicated as the most difficult item, FA3 (−2.52 logits) and FS4 (−2.55 logits) were the easiest items.



```
MEASURE    PERSON - MAP - ITEM
              <more>|<rare>
   5                +
                    |
                    |  NS8
                    |
                    |
                    |
   4                +
                 #  |
                    |  NS6
                    |
                    |T NS7
                    |  NS10
   3                +
                 .  |
                    |
                 .  |
                .#  |  NS5
                    |
   2       .###### T+  NS9
                    |
          .#######  |S
               .##  |  NA9
          .#######  |  FS9        NA6
            ###### S|  NA8
   1    .#########  +  NA7
     .############  |
      ############  |
     #############  |  FA4        FA5        NA10
     #############  |
     ############# M|  NA5
   0      ########  +M
        .#########  |  NS3
     #############  |  FS6        NS4
         .########  |  FS10       FS8
      .########### S|  FA10       FA7        FS7        NA3        NS2
            .#####  |  FA1        FS2        FS5        NA4
  -1          ####  +  FA9
             .####  |
                .#  |  FA8        NA2        NS1
               .## T|  FS1
              ####  |S FA2        FA6        NA1
   |            .#  |
  -2             .  +  FS3
                 #  |
                    |
                 .  |  FA3        FS4
                    |
                    |
  -3                +
                    |
                    |T
                    |
                    |
                    |
  -4                +
                    |
                    |
                    |
                    |
                    |
  -5                +
              <less>|<freq>
 EACH "#" IS 4: EACH "." IS 1 TO 3
```

Figure 19. Wright map for inductive reasoning test.

Reliability

The reliability criteria were evaluated following several indicators, including Rasch parameters using person and item reliability (Fisher, 2007; Linacre, 2021a), Cronbach's Alpha (α) (Taber, 2018) and McDonald's omega (ω) (Dunn et al., 2014). WINSTEPS software will generate person reliability, item reliability and Cronbach's Alpha (α), and SPSS was utilized to compute McDonald's omega (ω). Cronbach's Alpha (α) values ranged from 0.61 to 0.77 for all the tasks as well as the entire test, thus indicating sufficient reliability (Taber, 2018), and McDonald's omega (ω) ranges from 0.54 to 0.75, thus confirming acceptable reliability was achieved for only in the test level with 0.75 (Dunn et al., 2014). However, for person reliability and item reliability, the values range from 0.68 to 1.00. Fisher (2007) noted that values more than 0.67 demonstrated acceptable reliability. Overall, the adapted inductive reasoning test and all its tasks exhibited acceptable criteria for the Rasch reliability parameter. All the reliability results for both the tasks and test are summarized in Table 30.

Table 30. Reliability indicators

Reliability	Instrument				
	FA	FS	NA	NS	Test
Item reliability	0.99	0.99	0.99	1.00	1.00
Person reliability	0.69	0.68	0.68	0.71	0.79
Cronbach's Alpha (α)	0.70	0.61	0.65	0.62	0.77
McDonald's omega (ω)	0.66	0.58	0.57	0.54	0.75

DIF between paper-based and online-based tests

DIF analysis used in this study was the uniform DIF analysis that compares all ability levels of the two or more groups. DIF analysis based on the test method was performed to evaluate whether any item bias between paper-based and online-based tests was detected in student abilities. DIF analysis indicates participant responses based on subgroups for each item in the test (T. Bond & Fox, 2015; Boone et al., 2014; Khine, 2020). DIF can be assessed in accordance with two categories: a significant probability ($p < 0.05$) and DIF size. There are 3 DIF size classifications: negligible, slight to moderate ($| \text{DIF} | \geq 0.43$ logits), and moderate to large ($| \text{DIF} | \geq 0.64$ logits) (Zwick et al., 1999). The results of the DIF analysis revealed that 3 of the 40 items had a significant probability ($p < 0.05$), namely, FS2, FA7, and NS6. However, NS6 had moderate to large DIF. Furthermore, the online-based test was more difficult for students than the paper-based test with regard to NS6 item, with 0.94 logits of DIF size, p < 0.05. FS2 and FA7 were classified as having negligible DIF. The DIF analysis based on the test method is illustrated in Figure 20.

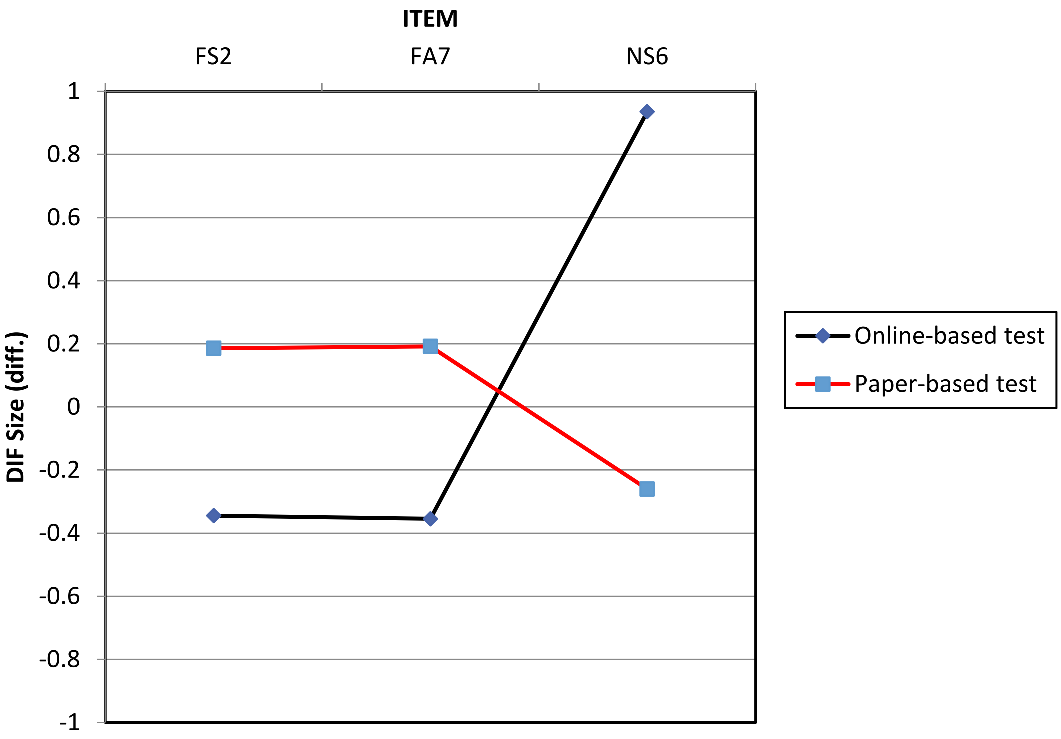


Figure 11. DIF analysis based on the test method

DIF across grade and gender

DIF analysis also is able to detect failures of invariance in this study context. As mentioned above in section 3.2, a significant probability (p < 0.05) and DIF size were used to identify DIF across grades and gender. Based on gender, four items have p < 0.05 with various DIF sizes, FS1(0.23 logits), NS2(0.34 logits), NS7(0.15 logits), NS8(0.16 logits) as present in figure. Therefore, we can categorize these four items as negligible DIF. DIF analysis is also performed based on grade level in Figure 22. Four items, FA1, NA4, NS3, NS5, have p < 0.05. However, these four items have DIF sizes below 0.43 logits. The highest DIF size between grade 10 and grade 12 has 0.42 logits which is still categorized as negligible DIF. We can assume the IR test can hold invariance confirming no DIF issue across grade and gender.

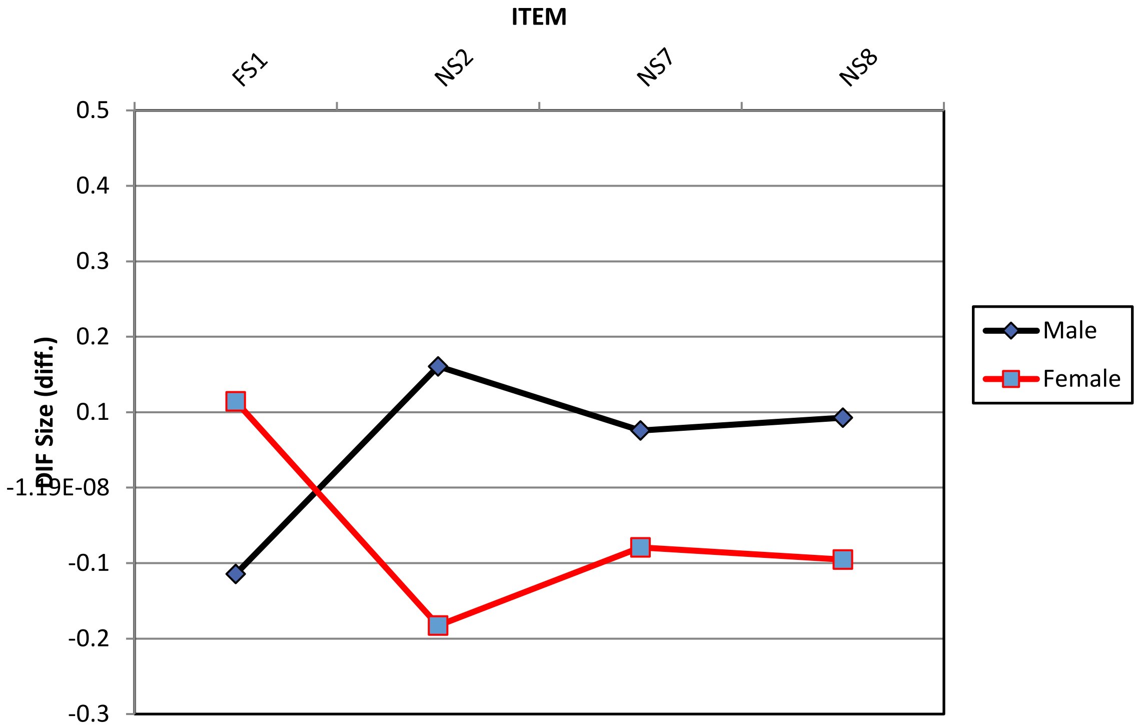


Figure 12. DIF analysis across gender

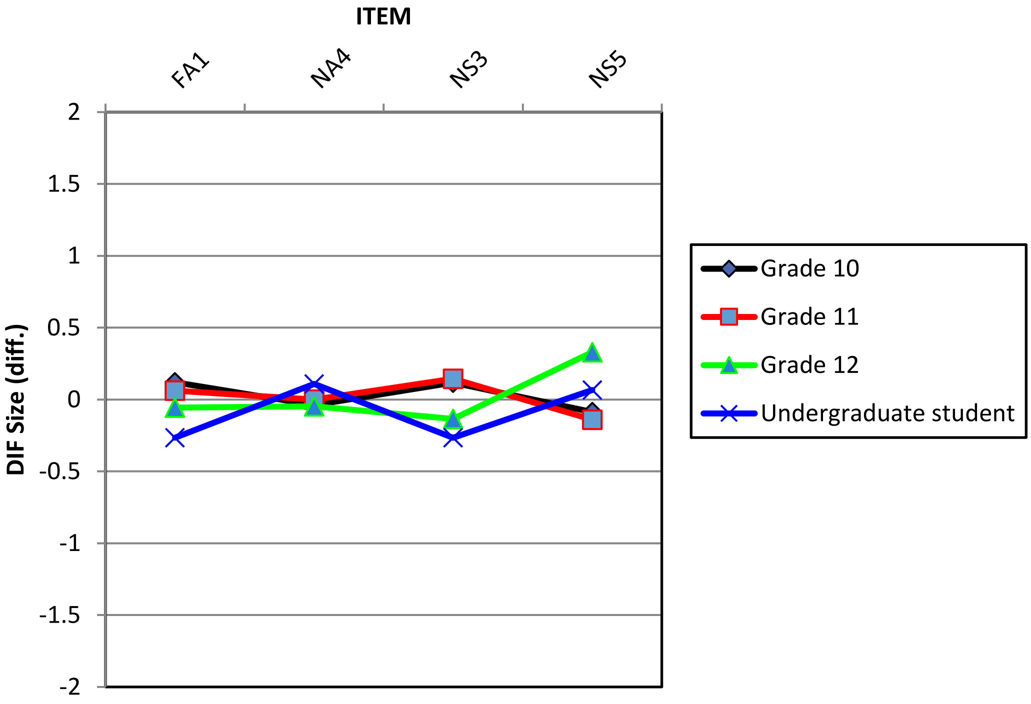


Figure 13. DIF analysis across grade

The evaluation of Indonesian students' inductive reasoning

The students' inductive reasoning based on the tasks of the test was evaluated. Rasch scales the students' ability from negative infinity to positive infinity whereby 0 logits is the average measure of students' ability (Bond & Fox, 2015; Sumintono & Widhiarso, 2014). In general, student abilities in solving the items of the inductive reasoning test were above average level (above 0 logits), with M = 0.24 logits; SD = 0.79. However, for the NA task (M = −0.04; SD = 0.78) and NS task (M = −1.41; SD = 0.98), the students' abilities were below average. These findings revealed that the students encountered some difficulties in solving numeric tasks, especially NS. This finding concurs with that of previous research on lower grades (Csapó, 1997; Díaz-Morales & Escribano, 2013; Molnár & Csapó, 2011; Pásztor et al., 2018; Sosa-Moguel & Aparicio-Landa, 2021; Van Vo & Csapó, 2020). The correlation matrix for all the tasks and the whole test were also evaluated. All correlation values were significant and ranged from 0.16 to 0.76. While the highest correlation was found between the FA task and inductive reasoning test (r = 0.76), the lowest correlation was revealed between the FS and NS tasks, even though the latter relationship was positively significant. This finding implied that students with a higher score on a task would achieve a higher score on the inductive reasoning test. The students' abilities and correlations between the inductive reasoning test and tasks are summarized in Table 31.

Table 25. Result of student abilities and correlation based on inductive reasoning test and tasks (N=40)

Test-subscale	M (logits)	SD	Logit range (Min, Max)	Pearson correlation			
				FA	FS	NA	NS
FA	1.16	0.8	(-2.58, 3.97)				
FS	0.98	1.01	(-2.72, 4.31)	.45**			
NA	-0.04	0.78	(-2.76, 4.05)	.36**	.24**		
NS	-1.41	0.98	(-4.25, 4.30)	.17**	.16**	.45**	
IR test	0.24	0.79	(-5.41, 1.69)	.76**	.68**	.74**	.56*

*Note. N = 856 *p < .05, **p < .001, M = Mean, SD = Standard deviation, IR = Inductive reasoning, FA = Figural analogies, FS = Figural series, NA = Number analogies, NS = Number series*

The students' inductive reasoning abilities were also evaluated in accordance with gender and grade. An examination of Table 32 reveals that undergraduate students outperformed students in other grades; M = 0.59; SD = 0.63. The 12th grade students had higher logit values (M = 0.31; SD = 0.66) than the 10th and 11th grade students.

Surprisingly, the 10th and 11th graders had the same logit values. Furthermore, the female students had superior performances (M = 0.28; SD = 0.88) in solving inductive reasoning problems in comparison to the male students. Person reliability for all the subgroups realized the minimum threshold criteria.

Table 26. Student inductive reasoning abilities based on gender and grade level (N=856).

	N	M (score)	M (logit)	SD	Person Reliability
Grade					
10	231	21.5	0.09	0.96	0.82
11	291	21.4	0.09	0.99	0.83
12	153	23	0.31	0.66	0.68
Undergraduate students	181	24.7	0.59	0.63	0.67
Gender					
Female	448	22.7	0.28	0.88	0.78
Male	408	22.1	0.19	0.9	0.79

To depict the primary trend between gender and grade related to the development of student inductive reasoning, graphical packages such as the yarrr package (Phillips, 2017) and the ggplot2 package (Wickham, 2016) were employed by using R software to create a pirate plot that combined the boxplot and student logit value distribution. An examination of Figure 23 reveals that males and females in each group were similar. The logit measure of females and males in all the groups remained stable, between 0 logits to 0.7 logits. No significant gender differences were identified in student inductive development. However, the distributions among the grades depicted different trends. While 10th and 11th grade students had similar inductive reasoning abilities, those in the 12th grade outperformed the former groups. The undergraduate students appeared to have the highest mean logit. However, further evaluation is needed to check any differences of students' inductive reasoning abilities among gender and grades.

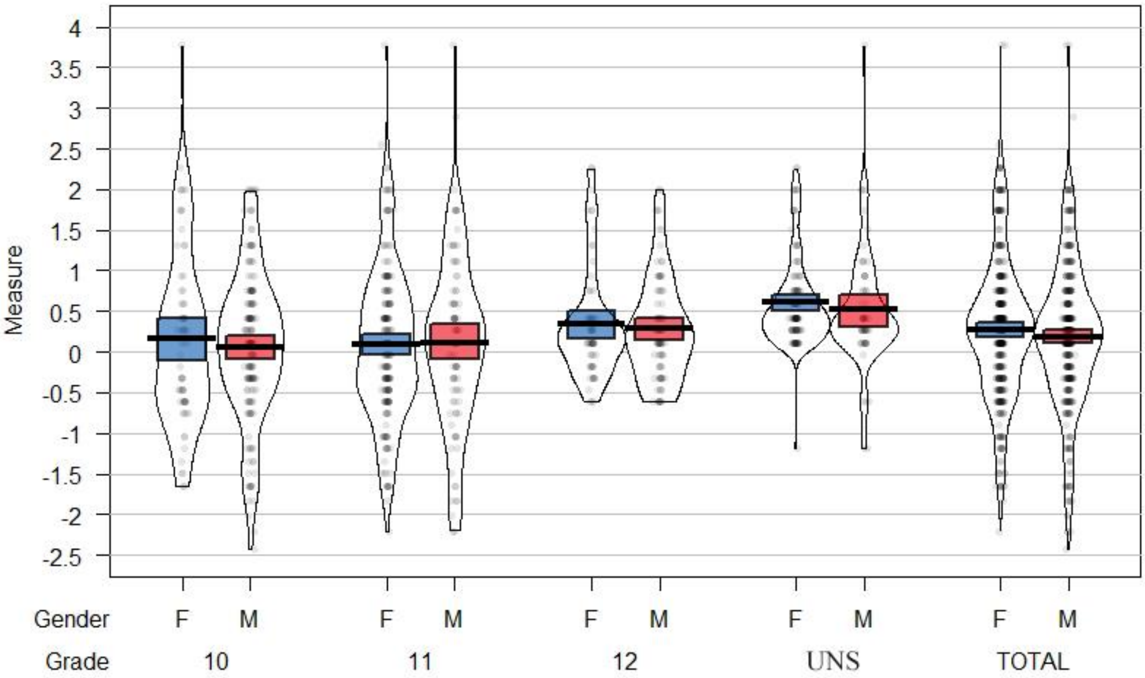


Figure 14. Pirate plot for comparing student measure (logit) based on gender and grade

An independent t-test was conducted to compare student inductive reasoning abilities between females and males for each grade. We classified the person logit values in accordance with grade and analyzed such by performing an independent t-test to determine gender differences. The results presented in Table 34 confirm that no significant differences were found between males and females for each grade. However, with the exception of the grade 11 students, the mean logit values of females were higher than those of males.

Table 27. The independent t-test for comparing student inductive reasoning abilities between females and males.

Grade	Female		Male		MD (logit)	t	p
	N	M/SD (logit)	N	M/SD (logit)			
10	68	0.17(1.08)	163	0.06(0.92)	0.11	0.74	.46
11	197	0.09(0.96)	94	0.11(1.08)	-0.02	-0.14	.89
12	63	0.35(0.69)	90	0.29(0.64)	0.06	0.11	.62
Undergraduate Student	120	0.62(0.56)	61	0.52(0.76)	0.10	0.91	.37

Moreover, a one-way ANOVA was employed using person logit values to check whether there were any differences among the grades. The results revealed significant differences among the grades for five groups that comprised four tasks and the inductive reasoning test: FA task [F (3, 855) = 16.35, p < 0.001], FS task [F (3, 855) = 12.00, p <

0.001], NA task [F (3, 855) = 4.36, p < 0.05], NS task [F (3, 855) = 6.33, p < 0.001], and entire inductive reasoning test [F (3, 855) = 15.01, p < 0.001]. To evaluate if there were any significant differences among the grades, WINSTEPS was employ to perform an independent t-test. The results are presented in Table 7. At the test level, significant differences were found among all the grades in the inductive reasoning test, with the exception of the 10th and 11th grade students. In relation to the tasks, the results demonstrated that the older students outperformed the younger students in all the tasks and the entire test. No significant difference was found between the 10th and 11th graders, thus demonstrating that these two groups of students had similar inductive reasoning abilities. Furthermore, the older students' performance was superior to that of their younger counterparts. Even though some significant differences were identified between tasks and the whole test, almost all mean differences revealed negative values, thus indicating that students in higher levels performed better than those in lower levels, except for the 10th and 11th grade students in the FA, NA, and NS tasks. The group comparisons for all the tasks and the entire test are presented in Table 34.

Table 28. The independent t-test for grade comparison based on IR test and tasks

Grade	IR test		FA		FS		NA		NS	
	MD (logit)	t	MD (logit)	t	MD (logit)	t	MD (logit)	t	MD (logit)	t
10th & 11th	-0.01	-0.04	0.01	0.05	-0.07	-0.45	0.03	0.23	0.1	0.81
10th & 12th	-0.22	-2.7**	-0.41	-3.23**	-0.5	-3.27**	-0.02	-0.02	-0.07	-0.51
10th & UNS	-0.5	-6.27**	-0.73	-6.09**	-0.83	-5.26**	-0.3	-2.32*	-0.43	-3.7**
11th & 12th	-0.22	-2.78**	-0.42	-3.49**	-0.43	-3**	-0.03	-0.23	-0.17	-1.28
11th & UNS	-0.49	-6.56**	-0.74	-6.6**	-0.76	-5.11**	-0.33	-2.62**	-0.53	-4.61**
12th & UNS	-0.27	-3.81**	-0.32	-2.68**	-0.33	-2.24*	-0.3	-2.1*	-0.36	-2.93**

*Note. N = 856 *p < .05, **p < .001, UNS = Undergraduate student, MD = Mean differences, IR = Inductive reasoning, FA = Figural analogies, FS = Figural series, NA = Number analogies, NS = Number series*

Item difficulties categorization for inductive reasoning test

In evaluating the difficulties of the inductive reasoning items, LVI analysis was used in accordance with the mean logit of items (0.00 logits), 1SD (1.69 logits), −1SD (−1.69 logits), the mean logit + 2SD (3.38), and the mean logit + 2SD (−3.38). The SD value demonstrated a wide dispersion of logit measures in item difficulty level. Using these thresholds, the inductive reasoning items were classified into 5 categories: difficulty level I (LVI ≥ mean logit + 2SD), difficulty level II (mean logit + 2SD > LVI ≥ 1SD), difficulty level III (1SD > LVI ≥ mean logit), difficulty level IV (mean logit > LVI ≥ −1SD), difficulty level V (LVI < −1SD) (Table 8). The results of the LVI analysis revealed 2 items (5%) in difficulty level I, 5 items (12.5%) in difficulty level II, 8 items (20%) in difficulty level III, 22 items in difficulty level IV, and 3 items (7.5%) in difficulty level V. These difficulty levels may be described as very difficult (difficulty level I), difficult (difficulty level II), moderate (difficulty level III), easy (difficulty level IV), and very easy (difficulty level V).

An examination of Table 8 reveals that two NS items (5%) were classified as very difficult even though four NS items were classified as easy items. One NA item and four NS items were classified as difficult items and two FA items, one FS item, and five NA items were classified as moderate. Finally, one FA item and two FS items were classified as very easy. This result corroborates with Figure 3 in relation to the Wright map, which conveys item and person scaling. In essence, items in the inductive reasoning test revealed a wide range of difficulty levels in relation to students' abilities. One can assume that the FA and FS items were easier than NS and NA items, which revealed that students' ability to solve figural tasks was more enhanced than their ability to solve numeric tasks. The classification in accordance with the LVI analysis is displayed in Table 35.

Table 29. The categorisation of inductive reasoning items difficulties

Task	Difficulty level I, LVI ≥ Mean logit + 2SD	Difficulty level II, Mean logit + 2SD > LVI ≥ 1SD	Difficulty level III, 1SD > LVI ≥ Mean logit	Difficulty level IV, Mean logit > LVI ≥ -1SD	Difficulty level V, LVI < -1SD
FA			FA4, FA5,	FA1, FA2, FA6, FA7, FA8, FA9, FA10	FA3
FS			FS9	FS1, FS2,FS5, FS6, FS7, FS8, FS10	FS3, FS4
NA		NA6	NA5,NA7, NA8, NA9, NA10	NA1, NA2, NA3, NA4	
NS	NS6, NS8	NS5, NS7, NS9, NS10		NS1, NS2, NS3, NS4	

Student inductive reasoning abilities categorization

LVP analysis was employed to classify students' ability to solve inductive reasoning problems in accordance with the mean logit of the person (0.24 logits), 1SD (0.89 logits), −1SD (−0.89 logits), the mean logit + 2SD (2.02 logits) and the mean logit + 2SD (−1.54 logits). LVP analysis was conducted in accordance with gender and grade by utilizing the COUNTIF function in Microsoft Excel to perform an automatic calculation of person logit measures. The results of LVP analysis, which resulted in four categories in relation to gender and grade, are presented in Table 36.

The results revealed that 10 (1.17%) females and 3 (0.35%) males were classified as having very high abilities. Furthermore, while 237 (27.02%) females were classified as possessing high abilities, 202 (23.60%) males had high abilities. Furthermore, 191 (22.31%) females and 184 (21.50%) males were classified as having moderate abilities, and 10 (1.17%) females and 19 (2.22%) males had low abilities. In relation to grade, 13 (1.52%) students were classified as having very high abilities, 6 of whom were in grade 12. In addition, 439 (51.29%) students possessed high abilities. Of these, 37% were undergraduate students. While 375 (43.81%) students were classified as having moderate ability, 29 students (3.39%) had low ability. Two (0.23%) 10th grade students were classified as having very high ability, 102 (11.92%) high ability, 114 (13.32%) moderate ability, and 13 (1.52%) low ability. In the 11th grade 6 (0.70%) students possessed very high ability, 122 (14.25%) high ability, 147 (17.17%) moderate ability, and 16 (1.87%) low ability. It is noteworthy that very few 12th graders and and undergraduate students possessed moderate ability and none in these groups were classified as having low ability. Rather, most were classified as having high abilities.

Table 30. The categorisation of student inductive reasoning abilities

Demographics	Very high, LVP > Mean Logit + 2SD	High, Mean Logit + 2SD ≥ LVP > Mean Logit	Moderate, Mean Logit ≥ LVP > Mean Logit - 2SD	Low, LVP < Mean Logit - 2SD
Gender				
Female	10	237	191	10
Male	3	202	184	19
Total	13	439	375	29
10th grade	2	102	114	13
11th grade	6	122	147	16
12th grade	2	74	77	0
Undergraduate student	3	141	37	0
Total	13	439	375	29

4.4.4 Discussion

The results revealed that the adapted inductive reasoning test for the Indonesian context is valid and reliable in accordance with Rasch parameters for measuring grade 10, 11, and 12 students at senior high school as well as undergraduate students. One may deduce that the inductive reasoning test can be employed among a wide range of grades in different cultural and country contexts. As noted previously, (Van Vo & Csapó, 2020) found an adaptation of the test was valid and reliable among 5th, 7th, 9th, and 11th graders in Vietnam. Furthermore, Csapó (1997) used a paper-based version of the inductive reasoning test to assess the inductive reasoning of 3rd to 11th grade students in Hungary.

Furthermore, DIF analysis based on the test method identified only one of the 40 items had a moderate to large DIF, thus implying that the items in the adapted version measured students' inductive reasoning, without the test method or media bias. This finding concurs with Csapó et al. (2009) who revealed no media bias was found when paper-based and online-based tests were compared. The application of technology through online-based testing was supported in this study because technology can offer several benefits for teachers, including developing item banks, using the adaptive test, composing anchoring tests, and collecting data continuously from a large sample. Csapó and Molnár (2019) proved that the media system successfully mapped more than 1,000 innovations (multimedia-supported) and operated in an experimental model in over 1,000 schools.

The evaluation of Indonesian students' inductive reasoning revealed that students tended to achieve higher performances in figural tasks with positive logits than numerical tasks. This finding concurs with previous studies (Feeney & Heit, 2007; Roberts et al., 2000; Van Vo & Csapó, 2020). The results further revealed that the females outperformed the males by 0.09 logits. However, no significant differences were found between females and males in all the grades (Table 6). These results concur with previous studies in Vietnam (Van Vo & Csapó, 2020), Namibia (Kambeyo & Wu, 2018), and Spain (Díaz-Morales & Escribano, 2013). In relation to grades, an independent t-test to compare grades also revealed significant differences among grades, with the exception of those in grade 10 and 11. Undergraduate students had higher inductive reasoning abilities than the other groups. These findings seem to support previous studies related to a comparison of student inductive reasoning among grades (Csapó 1997; Csapó et al. 2009, 2019; Díaz-Morales & Escribano, 2013; Van Vo & Csapó, 2020; Wu & Molnár, 2018). The evaluation of student inductive reasoning further revealed that the development of students' inductive development slowed after 14 years of age.

The classification of the difficulty of items showed that while most of the numeric items were classified as very difficult and difficult in accordance with LVI analysis, most figural tasks were classified as moderate, easy, and very easy. This finding is in line with Van Vo and Csapó’s (2020) evaluation of inductive reasoning that revealed students experienced difficulty solving numeric items: less than 40% of the answers were correct. The finding from

Van Vo and Csapó (2020) and Kambeyo & Wu (2018) also confirm that the figural tasks were relatively easy to solve compared to the numerics task. For instance, NS8 is the most difficult item in the numeric tasks based on Rasch scaling because this item requires complex pattern calculations with large numbers. Meanwhile, FS4 which is the easiest item in figural tasks only requires students to rotate the circle with a simple figure based on the previous pattern. LVP analysis revealed that students’ inductive reasoning abilities were classified into four categories. The results revealed that 439 (51,28%) were classified as having high abilities and 375 (43.8%) moderate abilities. This finding confirmed that students in higher grades could solve student inductive reasoning problems well. This is in line with previous studies (Csapó et al., 2019; Díaz-Morales & Escribano, 2013; Venville & Oliver, 2015).

This study contributed to the assessment of inductive reasoning using the Rasch measurement approach. The comprehensive analysis and application of the inductive reasoning assessment would extend the practical use of objective measurement in the educational field and encourage other researchers to explore inductive reasoning assessment in different contexts. Investigating person and item interactions based on individual-level statistics allowed the researcher to improve instrument quality and compared the result at the item level. This study also provided the item difficulty classification in the IR test. The person's ability measures was represented across grade and gender. Specific group comparison was also represented for four different tasks and the whole test. The DIF test performed in this study can examine the bias issue or failures of invariances in the assessment context.

Educators and teachers should be aware to identify student inductive reasoning that is related to their future academic achievement. The development of thinking skills in the learning process is embedded in the Indonesian core curriculum whereby inductive reasoning tests is also used as the primary test to examine student thinking skills in various job application. Additionally, Inductive reasoning can promote their cognitive development, develop their intelligence, and facilitate the understanding of the application of basic knowledge. Therefore, teachers and educators must understand the importance of inductive assessment and improve student inductive reasoning in the learning process.

4.4.5 Conclusion

In conclusion, the results of this study have contributed to an understanding of the item and person interaction on inductive reasoning. The adapted inductive reasoning test was shown to be valid and reliable in Indonesia and other countries, thus indicating this instrument can be employed in a wide range of cultural contexts. The items in the test are free of bias and only NS6 had a moderate to large DIF. Even though females outperformed males in relation to inductive reasoning abilities, no significant gender differences were found among the grades. Significant differences were found among all the groups, with the exception of the 10th and 11th grades. The classification of the difficulty of items revealed a wide range of difficulty levels, where numeric items were more difficult than figural items. Most of the students were classified as having high or moderate abilities.

The findings in this study provided initial information related to Indonesian students' inductive reasoning ability. This information may be useful for teachers and researchers to predict student success rates in other related subjects such as mathematics and science. In the Indonesian 2013 national curriculum, inductive reasoning is not included clearly. Accordingly, we believe that inductive reasoning skills can be embedded and trained in a wide range of grades because inductive reasoning is an essential thinking skill for predicting student academic achievement. We believe this study may be the first to perform different tests and utilize the Rasch measurement to assess students' inductive reasoning in Indonesia.

CHAPTER 5. CONCLUSIONS, RECOMMENDATIONS, LIMITATIONS

5.1 Conclusions

This dissertation includes two cross-sectional studies from pilot and main study with five published studies, one systematic literature review and four empirical study. In the first study in Chapter 2 is systematic review that was conducted on how often students have misconceptions about science was used to inform some findings. These findings included the different instruments used to find these misconceptions, the subjects on which students frequently have misconceptions, and the benefits and drawbacks of each test instrument. Some test instruments are used in combination to generate insightful results that can be used to support accurate interpretations of student misconceptions. Both written and oral tools have benefits and drawbacks. The technique of analysis can be strengthened by performing an integrated combination and by removing any flaws in a single instrument. Most researchers prefer simple multiple-choice tests (32.23%) and multiple tier tests (33.06%). According to study 1, researchers discovered that biology, chemistry, and physics subjects frequently lead to misconceptions among students. Biology had 15 concepts, chemistry had 12 concepts, and physics had 33 concepts. The systematic review provided evidence that the nature of misconception is resistant and tenacious to change, which poses a challenge for the advancement of scientific knowledge in the future. Those who wish to conduct study or teach with these tools must take great care to employ the appropriate techniques. Study 1 recommends three main steps before conducting research on misconceptions, including (1) examining the idea that typically causes misconceptions in students, (2) selecting a diagnostic tool based on benefits and drawbacks, (3) using combination two or more instrument to enhance research quality.

After conducting systematic literature review, the investigation of instrument validity and reliability was measure in first empirical study (Study 1) as pilot study, and study 2 as main study with larger sample size performed to invest item difficulty pattern. Student misconceptions in science evaluation is presented in Study 3. Pilot study in study 2 confirmed that all the items in the developed instrument are valid and reliable covering student ability based on item-person. The ANOVA test have verified that there are significant differences between science concepts across science disciplines and school grades whereby grade school predicted student misconception in science based on stepwise multiple regression. Independent sample t-test verified that no significant difference was found between boys and girls. Study 2 explores Evaluating item difficulty patterns for assessing student misconceptions in science across science subjects with larger sample size. Study 2 confirms that all items in the developed two-tier multiple choices diagnostic test meet the valid and reliable criteria. The item difficulty level of items on various science concepts is not universally based on science topics, but they are connected or similar across science disciplines, especially in physics, biology, and chemistry. Researchers also found items in the science concept may have different difficulty levels based on gender and grade. An empirical study of students' misconception in science was presented in Study 3. Study 3 confirmed significant differences in student conception mean scores between all cohorts; however, post-hoc analysis for ANOVA results evinced that differences were present only among 10th and 11th graders, and 10th and 12th graders in the biology subtest. In addition, the independent-sample t-test results confirmed that boys' and girls' mean scores were significantly different in that the former had higher mean scores than the latter, which demonstrated that boys tend to demonstrate better comprehension of science concepts and can solve science problems better than girls.

Lastly, Study 4 informed the findings in assessing student inductive reasoning comprehensively using Rasch measurement approach. The adapted inductive reasoning test was shown to be valid and reliable in Indonesia and other countries, thus indicating this instrument can be employed in a wide range of cultural contexts. The items in the test are free of bias and only NS6 had a moderate to large DIF. Even though females outperformed males in relation to inductive reasoning abilities, no significant gender differences were found among the grades. Significant differences were found among all the groups, with the exception of the 10th and 11th grades. The classification of the difficulty of items revealed a wide range of difficulty levels, where numeric items were more difficult than figural items. Most of the students were classified as having high or moderate abilities. in general, findings in this study provided initial information related to Indonesian students' inductive reasoning ability.

5.2 Educational implication

The findings from four different studies using cross/sectional approaches from pilot and main study have contributed to provide understanding and overview about misconceptions in science and inductive reasoning across gender and grade level. The instrument development and validation in this dissertation give insight to future researcher about how to investigate misconceptions in science and inductive reasoning from different viewpoints, such as investigating item difficulty pattern, student misconceptions, and performing validation with objective measurement.

The instrument development in form two-tier multiple choices and adapted inductive reasoning skills in this dissertation have achieved acceptable validity and reliability, so others research can use this instrument for their future research with different research questions. With valid and reliable instruments, researchers can save some effort and budget to perform further research. Therefore, the initial steps in performing instrument development and validation can have crucial implication to educational context.

The results have confirmed that Indonesian student have misconceptions in science in various science concepts. Students have difficulties in solving science problems in particular concepts because the concepts in science have different difficulty levels, such as Chemistry discipline having redox reaction as the most difficult concepts to solve. Based on person investigations. We found that not all students perform their knowledge precisely, 102 students have been indicated doing guessing in solving misconception in sciences. Therefore, teachers in learning activity can use the findings related to item difficulty patterns to prepare their lesson plan, and teachers can also do some investigation to see if there are guessing answers which perhaps indicate cheating activities.

By comparing different backgrounds, the findings from assessing misconceptions in science confirm that there is a bias in particular items. This indication can help educators or researchers to revise or omit items that have a bias especially by gender or group level. The result from comparing the mean based on gender and grade level show that there is no significant different between gender. Interestingly, we can confirm that even with different level of knowledge based on grade level, students still have not significant misconception values in science concepts which confirm that the misconception persistent and carried from student from lower to higher grade. These findings have crucial implications in educational context and confirm that misconceptions in science need to be treated properly to improve student success.

The adaptation inductive reasoning test has offered a valid and reliable instrument to be used by other researcher to measure inductive reasoning skill in Indonesia. Even though there are no special training for solving inductive reasoning problems, majority of Indonesian participants has been categorized into moderate and high category whereby the numerical items are more difficult than figural items. These findings are needed at least to give initial data and overview about student inductive reasoning in Indonesian context. Hence, the result from this research can be used as foundation to develop student inductive reasoning in Indonesian curriculum whereby the inductive reasoning test was often used for entrance test in higher education level and job carrier.

5.3 Recommendations

General recommendations based on series empirical studies in this dissertation presented as below:

1. Teachers or educators have to be aware of what kind of topics distributing misconceptions in science subject. Therefore, they can improve the student understanding about science concept and science achievement.
2. Screening for student understanding in the end of learning activity was needed using proper instrument, we recommended teachers can use the two-tier multiple choice diagnostics test to identify student knowledge and reasoning in a particular science concept.
3. For future researchers, pilot study as study 2 need to conducted before main study in study 3 and study 4 to confirm instrument validity and reliability in instrument development stage.
4. Future researchers can map the overall item difficulty level of whole science concepts.
5. Time series data collection or longitudinal research design must be added to explore whether there is a change of item difficulty level with the racking method in the Rasch measurement. Racking analysis allows researchers to evaluate whether there is a change in the difficulty level of the item on the different testing times sequentially.
6. The investigation of the relations between students' science misconceptions and thinking skills such as inductive reasoning and science reasoning is needed using the complex model using Structural Equation Modelling (SEM), not only assessing varaibles separately.
7. Future studies to mapping students' inductive reasoning needs to conduct using a longitudinal research design and include mixed methods.

5.4 Limitations

Some limitations based on series empirical studies in this dissertation presented as below:

1. Researchers did not develop items based on all scientific concepts studied in Indonesia. Items selected are based on concepts that distribute misconceptions in the previous research (AAAS, 2019; Allen, 2014; Csapó, 1998; Soeharto et al., 2019).
2. All respondents were from West Kalimantan, one of the provinces in Indonesia, one must exercise caution in generalizing the results to all Indonesian students, although the Rasch analysis have demonstrated that the samples hold local independence.
3. Studies in this dissertation performed quantitative analysis only; a mix of quantitative and qualitative methods may provide more meaningful insights.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and appreciation to my supervisor, Prof. Benő Csapó, for his unwavering guidance, support, and encouragement throughout the process of researching and writing this dissertation. His insightful feedback, constructive criticism, and valuable suggestions were instrumental in shaping the direction and focus of my work, and his patience and dedication to helping me improve my writing and analytical skills were truly invaluable. I am grateful for his friendship and mentorship which have undoubtedly enriched my academic experience and helped me to achieve my goals.

I also benefited significantly from Prof. Gyöngyvér Molnár, Prof. Edit Tóth, and Prof. Tibor Vígh, Prof. József Fejes Balázs experts that teach me how to conduct research with different statistical methods in their courses, Rasch measurement, Structural Equation Modelling, Fundamental statistics, and Advance statistics.

Some related institutions also support me to finish my study at University of Szeged. I truly thank Tempus Public Foundation that give me Stipendium Hungaricum Scholarship to pursue my doctoral degree. Ministry of Research and Education in Indonesia (DIKTI), STKIP Singkawang as my former institution from 2016-2021 before pursuing my doctoral study in Hungary, University of Vienna as a host for visiting researcher scholarship. Lastly, I would like to thank to the Hungarian Academy of Sciences (MTA) for giving me a funding to perform research stay for 1 year.

For my parents, I really appreciate their patience and support when I was in Hungary. I hope Allah blesses my parents plentifully. Lastly, with all my heart, I would like to thank my beloved wife Fitria Arifiyanti and my daughter Adaline Azkiya Arifiyanti for their support, patience, encouragement to me in finishing my study at the university of Szeged, Hungary

REFERENCES

AAAS. (2019). *Project 2061 | American Association for the Advancement of Science*. https://www.aaas.org/programs/project-2061

Abimbola, I. O. (1988). The problem of terminology in the study of student conceptions in science. *Science Education*, *72*(2), 175–184. https://doi.org/10.1002/sce.3730720206

Adadan, E., Savasci, F., & Martin, R. E. (2012). An analysis of 16–17-year-old students' understanding of solution chemistry concepts using a two-tier diagnostic instrument. *International Journal of Science Education*, *34*(4), 513–544.

Adams, D., Joo, M. T. H., Sumintono, B., & Oh, S. P. (2020). Blended Learning Engagement in Higher Education Institutions: A Differential Item Functioning Analysis of Students' Backgrounds. *Malaysian Journal of Learning and Instruction*, *17*(1), 133–158.

Allen, M. (2014). *Misconceptions in primary science*. McGraw-hill education.

Ancker, J., & Begg, M. (2017). Using Visual Analogies To Teach Introductory Statistical Concepts. *Numeracy*, *10*(2). https://doi.org/10.5038/1936-4660.10.2.7

Andrich, D. (2018). Advances in social measurement: A Rasch measurement theory. *Perceived Health and Adaptation in Chronic Disease*, 66–91.

Arslan, H. O., Cigdemoglu, C., & Moseley, C. (2012). A Three-Tier Diagnostic Test to Assess Pre-Service Teachers' Misconceptions about Global Warming, Greenhouse Effect, Ozone Layer Depletion, and Acid Rain. *International Journal of Science Education*, *34*(11), 1667–1686. https://doi.org/10.1080/09500693.2012.680618

Aryadoust, V., & Raquel, M. (2019). *Quantitative data analysis for language assessment volume I: Fundamental techniques*. Routledge.

Azizan, N. H., Mahmud, Z., & Rambli, A. (2020). Rasch Rating Scale Item Estimates using Maximum Likelihood Approach: Effects of Sample Size on the Accuracy and Bias of the Estimates. *International Journal of Advanced Science and Technology*, *29*(4), 2526–2531.

Badaruddin, K., & Hawi, A. (2022). Assessment of Student Attitudes in the 2013 Curriculum: Its Implementation and Problems. *Webology*, *19*(1), 6408–6419.

Bao, L., Cai, T., Koenig, K., Fang, K., Han, J., Wang, J., Liu, Q., Ding, L., Cui, L., & Luo, Y. (2009). Learning and scientific reasoning. *Science*, *323*(5914), 586–587. https://doi.org/10.1126/science.1167740

Barbic, S. P., & Cano, S. J. (2016). The application of Rasch measurement theory to psychiatric clinical outcomes research: Commentary on… Screening for depression in primary care. *BJPsych Bulletin*, *40*(5), 243–244.

Baweja, M. (2017). A study of errors and misconception in science in relation to scientific attitude among secondary school students. *International Journal of Advance Research*, *5*(3), 1707–1710. https://doi.org/10.21474/IJAR01/3682

Becker, N. M., & Cooper, M. M. (2014). College chemistry students' understanding of potential energy in the context of atomic-molecular interactions. *Journal of Research in Science Teaching*, *51*(6), 789–808. https://doi.org/10.1002/tea.21159

Bond, T., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 3rd Edition*. Routledge. https://doi.org/10.4324/9781315814698

Bond, T. G., & Fox, C. M. (2007). Rasch modeling applied: Rating scale design. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences (2nd Ed., Pp. 219–233). Mahwah, NJ: Lawrence Erlbaum Associates Publishers*.

Bond, T. G., Fox, C. M., & Lacey, H. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Boone, W. J. (2016). Rasch analysis for instrument development: Why,when,and how? *CBE Life Sciences Education*, *15*(4). https://doi.org/10.1187/cbe.16-04-0148

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer.

Boone, W. J., Townsend, J. S., & Staver, J. R. (2016). Utilizing multifaceted Rasch measurement through FACETS to evaluate science education data sets composed of judges, respondents, and rating scale items: An exemplar utilizing the elementary science teaching analysis matrix instrument. *Science Education*, *100*(2), 221–238.

Bradley, K. D., Peabody, M. R., Akers, K. S., & Knutson, N. (2015). Rating Scales in Survey Research: Using the Rasch model to illustrate the middle category measurement flaw. *Survey Practice, 8*(2).

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Bryman, A. (2016). *Social research methods*. Oxford university press.

Butler, J., Mooney Simmie, G., & O'Grady, A. (2015). An investigation into the prevalence of ecological misconceptions in upper secondary students and implications for pre-service teacher education. *European Journal of Teacher Education*, *38*(3), 300–319. https://doi.org/10.1080/02619768.2014.943394

Caleon, I., & Subramaniam, R. (2010). Development and application of a three-tier diagnostic test to assess secondary students' understanding of waves. *International Journal of Science Education*, *32*(7), 939–961. https://doi.org/10.1080/09500690902890130

Chabalengula, Mweene, V., & Sanders, M. (2012). Diagnosing Students ' Understanding of Energy and. *International Journal of Science and Mathematics Education*, *10*(June 2010), 241–266.

Chan, S.-W., Looi, C.-K., & Sumintono, B. (2021). Assessing computational thinking abilities among Singapore secondary students: A Rasch model measurement analysis. *Journal of Computers in Education*, *8*(2), 213–236. https://doi.org/10.1007/s40692-020-00177-2

Chazbeck, B., & Ayoubi, Z. (2018). Resources Used by Lebanese Secondary Physics Teachers' for Teaching Electricity: Types, Objectives and Factors Affecting their Selection. *Journal of Education in Science, Environment and Health*, *4*(2), 118–128. https://doi.org/10.21891/jeseh.409487

Chen, R. F., Scheff, A., Fields, E., Pelletier, P., & Faux, R. (2014). Mapping energy in the Boston public schools curriculum. In *Teaching and Learning of Energy in K–12 Education* (pp. 135–152). Springer.

Childers, J. B., & Exemplars, M. (2020). Language and Concept Acquisition from Infancy Through Childhood. In *Language and Concept Acquisition from Infancy Through Childhood*. https://doi.org/10.1007/978-3-030-35594-4

Chou, Y.-T., & Wang, W.-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement, 70*(5), 717–731.

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q 3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement*, *41*(3), 178–194.

Chu, S. K. W., Reynolds, R. B., Tavares, N. J., Notari, M., & Lee, C. W. Y. (2017). Twenty-First Century Skills and Global Education Roadmaps. In *21st Century Skills Development Through Inquiry-Based Learning* (pp. 17–32). Springer Singapore. https://doi.org/10.1007/978-981-10-2481-8_2

Cohen, L., Lawrence, M., & Morrison, K. (2018). Research Methods in Education. Eighth Edition. In *Research Methods in Education*. Routledge. https://doi.org/10.1111/j.1467-8527.2007.00388_4.x

Cooper, M. M., & Klymkowsky, M. W. (2013). The trouble with chemical energy: Why understanding bond energies requires an interdisciplinary systems approach. *CBE—Life Sciences Education*, *12*(2), 306–312.

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications.

Csapó, B. (1997). The Development of Inductive Reasoning: Cross-sectional Assessments in an Educational Context. *International Journal of Behavioral Development*, *20*(4), 609–626. https://doi.org/10.1080/016502597385081

Csapó, B. (1998). *Iskolai tudas*. Osiris Kiadó.

Csapó, B. (2012). Developing a framework for diagnostic assessment of early science. *S Bernholt, S., Neumann, K. És Nentwig, P.(Szerk.): Making It Tangible–Learning Outcomes in Science Education. Waxmann, Münster*, 55–78.

Csapó, B., Hotulainen, R., Pásztor, A., & Molnár, G. (2019). Az induktív gondolkodás fejlődésének összehasonlító vizsgálata:online felmérések Magyarországon és Finnországban. *Neveléstudomány*, *2019*(3–4), 5–24. https://doi.org/10.21549/NTNY.27.2019.3.1

Csapó, B., & Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in Psychology*, *10*(JULY). https://doi.org/10.3389/fpsyg.2019.01522

Csapó, B., Molnár, G., & Tóth, K. R. (2009). Comparing paper-and-pencil and online assessment of reasoning skills: A pilot study for introducing eletronic testing in large-scale assessment in Hungary. In *The transition to computer-based assessment: New approaches to skills assessment and implications for large-*

scale testing (Issue 2). Luxemburg: Office for Official Publications of the European Communities.

Csapó, B., & Szabó, G. (Eds.). (2012). *Framework for diagnostic assessment of science* (First edition). Nemzeti Tankönyvkiadó.

De Koning, E., Hamers, J. H. M., Sijtsma, K., & Vermeer, A. (2002). Teaching inductive reasoning in primary education. *Developmental Review*, *22*(2), 211–241. https://doi.org/10.1006/drev.2002.0548

Díaz-Morales, J. F., & Escribano, C. (2013). Predicting school achievement: The role of inductive reasoning, sleep length and morningness–eveningness. *Personality and Individual Differences*, *55*(2), 106–111. https://doi.org/10.1016/j.paid.2013.02.011

Driver, R., & Easley, J. (1978). Pupils and Paradigms: A Review of Literature Related to Concept Development in Adolescent Science Students. *Studies in Science Education*, *5*(1), 61–84. https://doi.org/10.1080/03057267808559857

Duit, R. (2014). Teaching and Learning the Physics Energy Concept. In *Teaching and Learning of Energy in K – 12 Education* (pp. 67–85). Springer International Publishing. https://doi.org/10.1007/978-3-319-05017-1_5

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*, 399–412. https://doi.org/10.1111/bjop.12046

Eggen, P. D., Kauchak, D. P., & Garry, S. (2007). *Educational psychology: Windows on classrooms*. Pearson/Merrill/Prentice Hall.

Engelhard Jr, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Erman, E. (2017). Factors contributing to students' misconceptions in learning covalent bonds. *Journal of Research in Science Teaching*, *54*(4), 520–537. https://doi.org/10.1002/tea.21375

Eshach, H., Lin, T., & Tsai, C. (2018). Misconception of sound and conceptual change: A cross-sectional study on students' materialistic thinking of sound. *Journal of Research in Science Teaching*, *55*(5), 664–684.

Faisal, & Martin, S. N. (2019). Science education in Indonesia: Past, present, and future. *Asia-Pacific Science Education*, *5*(1). https://doi.org/10.1186/s41029-019-0032-0

Fajarini, F., Utari, S., & Prima, E. C. (2018). Identification of students' misconception against global warming concept. *International Conference on Mathematics and Science Education of Universitas Pendidikan Indonesia, 3*, 199–204.

Fariyani, Q., Rusilowati, A., & Sugianto, S. (2017). Four-tier diagnostic test to identify misconceptions in geometrical optics. *Unnes Science Education Journal*, *6*(3).

Feeney, A., & Heit, E. (2007). *Inductive Reasoning Experimental, Developmental, and Computational Approaches*. NY:Cambridge University Press.

Fisher, W. P. J. (2007). Rating Scale Instrument Quality Criteria. *Rasch Measurement Transactions, 21*(1), 1095.

Fotou, N., & Abrahams, I. (2020). Extending the Role of Analogies in the Teaching of Physics. *The Physics Teacher*, *58*(1), 32–34. https://doi.org/10.1119/1.5141968

Fuentes, P. L. S. (2021). Mathematical Ability, Level of Science Misconceptions, and Science Performance of First-Year College Students. *International Journal of Advanced Engineering, Management and Science, 7*, 3. https://doi.org/10.22161/ijaems.73.4

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). Educational research: An introduction. *AE Burvikovs, Red.) USA: Pearson*.

Galvin, E., & Mooney, S. G. (2015). Identification of Misconceptions in the Teaching of Biology: A Pedagogical Cycle of Recognition, Reduction and Removal. *Higher Education of Social Science*, *8*(2), 1–8. https://doi.org/10.3968/6519

Greiff, S., Molnár, G., Martin, R., Zimmermann, J., & Csapó, B. (2018). Students' exploration strategies in computer-simulated complex problem environments: A latent class approach. *Computers & Education*, *126*, 248–263.

Griffin, P. (2010). *Item response modelling: An introduction to the Rasch model.* Assessment Research Centre Faculty of Education, The University of Melbourne.

Gurbuz, F. (2015). Physics Education: Effect of Micro-teaching Method Supported by Educational Technologies on Pre-service Science Teachers' Misconceptions on Basic Astronomy Subjects. *Journal of Education and Training Studies*, *4*(2), 2010–2011. https://doi.org/10.11114/jets.v4i2.1140

Gurcay, D., & Gulbas, E. (2015). Development of three-tier heat, temperature, and internal energy diagnostic test. *Research in Science and Technological Education*, *33*(2), 197–217. https://doi.org/10.1080/02635143.2015.1018154

Gurel, D. K., Eryilmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(5), 989–1008. https://doi.org/10.12973/eurasia.2015.1369a

Gyöngyvér, M., & Csapó, B. (2011). *Az 1–11 évfolyamot átfogó induktív gondolkodás kompetenciaskála készítése a valószínűségi tesztelmélet alkalmazásával [Constructing inductive reasoning competency scales for years 1–11 using IRT models]*. *111*(2), 127–140.

Hagell, P. (2014). Testing rating scale unidimensionality using the principal component analysis (PCA)/t-test protocol with the Rasch model: The primacy of theory over statistics. *Open Journal of Statistics*, *4*(6), 456–465.

Harman, G., & Çökelez, A. (2017). Analojilerin Fen Eğitimindeki Yeri ve Önemi. *Necatibey Eğitim Fakültesi Elektronik Fen ve Matematik Eğitimi Dergisi*, *11*(1), Article 1. https://doi.org/10.17522/balikesirnef.356303

Hasan, S., Bagayoko, D., & Kelley, E. L. (1999). Misconceptions and the certainty of response index (CRI). *Physics Education*, *34*(5), 294–299.

Hasan, S. H. (2013a). History Education in Curriculum 2013: A New Approach To Teaching History. *Historia: Jurnal Pendidik Dan Peneliti Sejarah*, *14*(1), 163–178. https://doi.org/10.17509/historia.v14i1.2023

Hasan, S. H. (2013b). History Education in Curriculum 2013: A New Approach To Teaching History. *Historia: Jurnal Pendidik Dan Peneliti Sejarah*, *14*(1), 163. https://doi.org/10.17509/historia.v14i1.2023

Hayes, B. K., & Heit, E. (2017). Inductive reasoning 2.0. *Wiley Interdisciplinary Reviews: Cognitive Science*, *9*(3), 1–13. https://doi.org/10.1002/wcs.1459

Hesti, R. (2021). Analogy Educational Comics to Overcome Students' Misconception on Simple Electricity Circuit Material. *Journal of Physics: Conference Series*, *1957*(1), 012036. https://doi.org/10.1088/1742-6596/1957/1/012036

Hotulainen, R., Pásztor, A., Kupiainen, S., Molnár, G., & Csapó, B. (2018). Entering school with equal skills? A two-country comparison of early inductive reasoning. *August Paper Presented at the 9th Biennial Conference of EARLI*

SIG 1: Assessment and Evaluation: Assessment & Learning Analytics, Paper: C_2_3.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

IBM Corp. (2017). *IBM SPSS Statistics for Windows (Version 25) [Computer software]*. Armonk, NY: IBM SPSS Corp.

Istikomah, F., Rochmad, R., & Winarti, E. R. (2017). Analysis of 7th Grade Students' Inductive Reasoning Skill in PBL-Bertema Model Towards Responsibility Character. *Unnes Journal of Mathematics Education*, *6*(3), 345–351. https://doi.org/10.15294/ujme.v6i3.17600

Kaltakci-Gurel, D., Eryilmaz, A., & McDermott, L. C. (2017). Development and application of a four-tier test to assess pre-service physics teachers' misconceptions about geometrical optics. *Research in Science and Technological Education*, *35*(2), 238–260. https://doi.org/10.1080/02635143.2017.1310094

Kambeyo, L., & Wu, H. (2018). Online assessment of students' inductive reasoning skills abilities in Oshana Region, Namibia. *International Journal of Educational Sciences*, *21*(1), 1–12. https://doi.org/11. 258359/KRE-86

Keeley, P. (2012). Misunderstanding misconceptions. *Science Scope*, *35*(8), 12–13.

Kerr, K., Beggs, J., & Murphy, C. (2006). Comparing children's and student teachers' ideas about science concepts. *Irish Educational Studies*, *25*(3), 289–302. https://doi.org/10.1080/03323310600913732

Khine, M. S. (2020). Rasch Measurement. In *Rasch Measurement*. https://doi.org/10.1007/978-981-15-1800-3

Kinshuk, Lin, T., & Mcnab, P. (2006). Cognitive trait modelling: The case of inductive reasoning ability. *Innovations in Education and Teaching International*, *43*(2), 151–161. https://doi.org/10.1080/14703290600650442

Kiray, S. A., Aktan, F., Kaynar, H., Kilinc, S., & Gorkemli, T. (2015). A descriptive study of pre-service science teachers' misconceptions about sinking-floating. *Asia-Pacific Forum on Science Learning & Teaching*, *16*(2), 1–28.

Kiray, S. A., & Simsek, S. (2021). Determination and Evaluation of the Science Teacher Candidates' Misconceptions About Density by Using Four-Tier

Diagnostic Test. *International Journal of Science and Mathematics Education, 19*(5), 935–955. https://doi.org/10.1007/s10763-020-10087-5

Kirbulut, Z. D., & Geban, O. (2014). Using three-tier diagnostic test to assess students' misconceptions of states of matter. *Eurasia Journal of Mathematics, Science and Technology Education*, *10*(5), 509–521. https://doi.org/10.12973/eurasia.2014.1128a

Klauer, K. J. (1996). Teaching inductive reasoning: Some theory and three experimental studies. *Learning and Instruction*, *6*(1), 37–57. https://doi.org/10.1016/S0959-4752(96)80003-X

Klauer, K. J., & Phye, G. D. (2008). Inductive Reasoning: A Training Approach. *Review of Educational Research, 78*(1), 85–123. https://doi.org/10.3102/0034654307313402

Kleppang, A. L., Steigen, A. M., & Finbråten, H. S. (2020). Using Rasch measurement theory to assess the psychometric properties of a depressive symptoms scale in Norwegian adolescents. *Health and Quality of Life Outcomes*, *18*(1), 1–8. https://doi.org/10.1186/s12955-020-01373-5

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Korkmaz, S. D., Bahadir, A., Aybek, E. C., & Suat, P. A. T. (2018). Evaluating the gifted students' understanding related to plasma state using plasma experimental system and two-tier diagnostic test. *Journal of Education in Science Environment and Health*, *4*(1), 46–53.

Korom, E., Németh, M., Pásztor, A., & Csapó, B. (2017). Relationship between scientific and inductive reasoning in grades 5 and 7. *17thBiennial Conference of the European Association for Research on Learning and Instruction (EARLI)*, 128–129.

Korur, F. (2015). Exploring seventh-grade students' and pre-service science teachers' misconceptions in astronomical concepts. *Eurasia Journal of Mathematics, Science and Technology Education*, *11*(5), 1041–1060. https://doi.org/10.12973/eurasia.2015.1373a

Köse, S. (2004). Effectiveness of conceptual change texts accompanied with concept mapping instructions on overcoming prospective science teachers' misconceptions of photosynthesis and respiration in plants. *Published Ph. D.,* *Karadeniz Technical University, Institute of Natural and Applied Sciences, Trabzon*.

Krajcik, J., Chen, R. F., Eisenkraft, A., Fortus, D., Neumann, K., Nordine, J., & Scheff, A. (2014). Conclusion and summary comments: Teaching energy and associated research efforts. In *Teaching and Learning of Energy in K–12 Education* (pp. 357–363). Springer.

Lafraire, J., Rioux, C., Hamaoui, J., Girgis, H., Nguyen, S., & Thibaut, J. P. (2020). Food as a borderline domain of knowledge: The development of domain-specific inductive reasoning strategies in young children. *Cognitive Development, 56*(June), 100946. https://doi.org/10.1016/j.cogdev.2020.100946

Laliyo, L. A. R., Botutihe, D. N., & Panigoro, C. (2019). The Development of Two-Tier Instrument Based On Distractor to Assess Conceptual Understanding Level and Student Misconceptions in Explaining Redox Reactions. *International Journal of Learning, Teaching and Educational Research*, *18*(9), 216–237. https://doi.org/10.26803/ijlter.18.9.12

Laliyo, L. A. R., Puluhulawa, F. U., Eraku, S., & Salimi, Y. K. (2020). The Prevalence of Students and Teachers' Ideas about Global Warming and the Use of Renewable Energy Technology. *Journal of Environmental Accounting and Management*, *8*(3), 243–256. https://doi.org/10.5890/jeam.2020.09.003

Lancor, R. (2015). An analysis of metaphors used by students to describe energy in an interdisciplinary general science course. *International Journal of Science Education, 37*(5–6), 876–902.

Lancor, R. A. (2014). Using student-generated analogies to investigate conceptions of energy: A multidisciplinary study. *International Journal of Science Education, 36*(1), 1–23.

Leaper, C., Farkas, T., & Brown, C. S. (2012). Adolescent girls' experiences and gender-related beliefs in relation to their motivation in math/science and English. *Journal of Youth and Adolescence, 41*(3), 268–282. https://doi.org/10.1007/s10964-011-9693-z

Lee, Y., Kim, M. L., & Hong, S. (2021). Big-data analytics: Exploring the well-being trend in South Korea through inductive reasoning. *KSII Transactions on Internet and Information Systems, 15*(6), 1–16. https://doi.org/10.3837/tiis.2021.06.003

Leedy, P. D., & Ormrod, J. E. (2005). *Practical research: Planning and design* (Vol. 1). Pearson.

Leighton, J. P., & Sternberg, R. J. (2003). *The Nature of Reasoning* (J. P. Leighton & R. J. Sternberg, Eds.). Cambridge University Press. https://doi.org/10.1017/CBO9780511818714

Liampa, V., Malandrakis, G. N., Papadopoulou, P., & Pnevmatikos, D. (2019). Development and Evaluation of a Three-Tier Diagnostic Test to Assess Undergraduate Primary Teachers' Understanding of Ecological Footprint. *Research in Science Education*, *49*(3), 711–736. https://doi.org/10.1007/s11165-017-9643-1

Lin, T., & McNab, P. (2006). Adaptive support for inductive reasoning ability. In *Web-Based Intelligent E-Learning Systems: Technologies and Applications* (pp. 1–23). IGI Global.

Linacre, John M. (2021a). *Winsteps® Rasch measurement computer program User's Guide*. Winsteps.com.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, *2*, 266–283.

Linacre, J. M. (2020). *Winsteps® (Version 4.7.0) [Computer Software].* (4.7.0). Winsteps.com.

Liu, O. L., Ryoo, K., Linn, M. C., Sato, E., & Svihla, V. (2015). Measuring Knowledge Integration Learning of Energy Topics: A two-year longitudinal study. *International Journal of Science Education*, *37*(7), 1044–1066. https://doi.org/10.1080/09500693.2015.1016470

Liu, X. (2007). Elementary to high school students' growth over an academic year in understanding concepts of matter. *Journal of Chemical Education*, *84*(11), 1853.

Martin, R. E. (2005). *Teaching science for all children: Inquiry methods for constructing understanding*. Allyn & Bacon.

Maryanto, A. (2019). The Effectiveness of Inductive and Deductive Strategies to Improve Motivation and Achievement in Learning Science of Junior High School Students. *Journal of Science Education Research*, *3*(1), 1–10. https://doi.org/10.21831/jser.v3i1.27296

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (2005). Assessing Science Understanding. In *Assessing Science Understanding*. Elsevier. https://doi.org/10.1016/B978-0-12-498365-6.X5000-8

Molnár, G., Greiff, S., & Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: Relations and development. *Thinking Skills and Creativity, 9*, 35–45. https://doi.org/10.1016/j.tsc.2013.03.002

Morrison, G. R., Ross, S. J., Morrison, J. R., & Kalman, H. K. (2019). *Designing effective instruction*. John Wiley & Sons.

Mubarokah, F. D., Mulyani, S., & Indriyanti, N. Y. (2018). Identifying students' misconceptions of acid-base concepts using a three-tier diagnostic test: A case of Indonesia and Thailand. *Journal of Turkish Science Education*, *15*(Special Issue), 51–58. https://doi.org/10.12973/tused.10256a

Murdoch, J. (2018). Our preconceived notions of play need to challenging. *Early Years Educator, 19*(9), 22–24. https://doi.org/10.12968/eyed.2018.19.9.22

Muthén, B. O. (1993). Goodness of fit with categorical and other nonnormal variables. In *SAGE Focus Editions* (Vol. 154, p. 205). Sage Publications.

Muthén, L., & Muthén, B. (2017). *Mplus user's guide (eight edition)[Computer software manual]*. Los Angeles, CA: Muthén & Muthén.

Natsir, Y., Qismullah Yusuf, Y., & Fiolina Nasution, U. (2018). The Rise and Fall of Curriculum 2013: Insights on the Attitude Assessment from Practicing Teachers. *SHS Web of Conferences*, *42*, 00010. https://doi.org/10.1051/shsconf/20184200010

Neumann, K., Viering, T., Boone, W. J., & Fischer, H. E. (2013). *Towards a Learning Progression of Energy*. *50*(2), 162–188. https://doi.org/10.1002/tea.21061

Nikolov, M., & Csapó, B. (2018). The relationships between 8th graders' L1 and L2 reading skills, inductive reasoning and socio-economic status in early English and German as a foreign language programs. *System*, *73*, 48–57. https://doi.org/10.1016/j.system.2017.11.001

OECD. (2016). *PISA 2015 results (Volume I): Excellence and equity in education*.

OECD. (2020). *Science performance (PISA) (indicator)*. OECD. https://doi.org/doi: 10.1787/91952204-en

Park, M., & Liu, X. (2019). An Investigation of Item Difficulties in Energy Aspects Across Biology, Chemistry, Environmental Science, and Physics. *Research in Science Education*. https://doi.org/10.1007/s11165-019-9819-y

Pásztor, A. (2016). *Technology-based assessment and development of inductive reasoning*. 10.14232/phd.3191

Pásztor, A., Kupiainen, S., Hotulainen, R., Molnár, G., & Csapó, B. (2018). Comparing Finnish and Hungarian fourth grade students' inductive reasoning skills. *9th Biennial Conference of EARLI SIG 1: Assessment and Evaluation: Assessment & Learning Analytics.*, *1*, Paper: A_1_3.

Perret, P. (2015). Children's Inductive Reasoning: Developmental and Educational Perspectives. *Journal of Cognitive Education and Psychology*, *14*(3), 389–408. https://doi.org/10.1097/NNR.0b013e31824798ba

Peşman, H., & Eryılmaz, A. (2010). Development of a three-tier test to assess misconceptions about simple electric circuits. *The Journal of Educational Research*, *103*(3), 208–222. https://doi.org/10.1080/00220670903383002

Peterson, R. F., Treagust, D. F., & Garnett, P. (1989). Development and application of a diagnostic instrument to evaluate grade-11 and-12 students' concepts of covalent bonding and structure following a course of instruction. *Journal of Research in Science Teaching*, *26*(4), 301–314.

Phillips, N. D. (2017). Yarrr! The pirate's guide to R. In *APS Observer* (Vol. 30, Issue 3).

Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2), 1–14. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Prastowo, A., & Fitriyaningsih, F. (2020). Learning Material Changes as the Impact of the 2013 Curriculum Policy for the Primary School/Madrasah Ibtidaiyah. *Edukasia : Jurnal Penelitian Pendidikan Islam*, *15*(2), 251. https://doi.org/10.21043/edukasia.v15i2.7947

Prodjosantoso, A. K., Hertina, A. M., & Irwanto. (2019). The misconception diagnosis on ionic and covalent bonds concepts with three tier diagnostic test. *International Journal of Instruction*, *12*(1), 1477–1488. https://doi.org/10.29333/iji.2019.12194a

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ramirez, A., Sarathy, S. M., & Gascon, J. (2020). CO2 Derived E-Fuels: Research Trends, Misconceptions, and Future Directions. *Trends in Chemistry*, *2*(9), 785–795. https://doi.org/10.1016/j.trechm.2020.07.005

Ratnasari, D., & Suparmi, S. (2017). Effect of problem type toward students' conceptual understanding level on heat and temperature. *Journal of Physics: Conference Series*, *909*(1), 12054.

Roberts, M. J., Welfare, H., Livermore, D. P., & Theadom, A. M. (2000). Context, visual salience, and inductive reasoning. *Thinking & Reasoning*, *6*(4), 349–374. https://doi.org/10.1080/135467800750038175

Samsudin, A., Afif, N. F., Nugraha, M. G., Suhandi, A., Fratiwi, N. J., Aminudin, A. H., Adimayuda, R., Linuwih, S., & Costu, B. (2021). Reconstructing Students' Misconceptions on Work and Energy through the PDEODE* E Tasks with Think-Pair-Share. *Journal of Turkish Science Education*, *18*(1), 118–144. https://doi.org/10.36681/tused.2021.56

Seo, E., Shen, Y., & Alfaro, E. C. (2019). Adolescents' beliefs about math ability and their relations to STEM career attainment: Joint consideration of race/ethnicity and gender. *Journal of Youth and Adolescence, 48,* 306–325. https://doi.org/10.1007/s10964-018-0911-9

Sholihah, R., Siswanto, J., Roshayanti, F., & Nugroho, A. S. (2021). Profil analogical reasoning siswa sma. *INKUIRI: Jurnal Pendidikan IPA*, *10*(1), Article 1. https://doi.org/10.20961/inkuiri.v10i1.39484

Siswono, T. Y. E., Hartono, S., & Kohar, A. W. (2020). Deductive or Inductive? Prospective Teachers' Preference of Proof Method on An Intermediate Proof Task. *Journal on Mathematics Education*, *11*(3), 417–438. https://doi.org/10.22342/jme.11.3.11846.417-438

Slater, E. V., Morris, J. E., & McKinnon, D. (2018). Astronomy alternative conceptions in pre-adolescent students in Western Australia. *International Journal of Science Education*, *40*(17), 2158–2180. https://doi.org/10.1080/09500693.2018.1522014

Soeharto, Csapó, B., Sarimanah, E., Dewi, F. I., & Sabri, T. (2019). A review of students' common misconceptions in science and their diagnostic assessment tools. *Jurnal Pendidikan IPA Indonesia*, *8*(2), 247–266. https://doi.org/10.15294/jpii.v8i2.18649

Soeharto, S. (2017). PHYCCTM Development Based On KKNI On Impuls And Momentum Material To Increase HOTS And Independent Character. *EDUCATIO : Journal of Education*, *2*(2). https://doi.org/10.29138/educatio.v2i2.187

Soeharto, S. (2021). Evaluation and development of students' misconception using diagnostic assessment in science across school grades: A Rasch measurement approach. *Journal of Turkish Science Education*, *18*, 351–370. https://doi.org/10.36681/tused.2021.78

Soeharto, S., & Csapó, B. (2021). Evaluating item difficulty patterns for assessing student misconceptions in science across physics, chemistry, and biology concepts. *Heliyon*, *7*(11), e08352. https://doi.org/10.1016/j.heliyon.2021.e08352

Soeharto, S., & Csapó, B. (2022a). Assessing Indonesian student inductive reasoning: Rasch analysis. *Thinking Skills and Creativity*, *46*, 1–16. https://doi.org/10.1016/j.tsc.2022.101132

Soeharto, S., & Csapó, B. (2022b). Exploring Indonesian student misconceptions in science concepts. *Heliyon*, *8*(9), e10720. https://doi.org/10.1016/j.heliyon.2022.e10720

Soeharto, S., Csapő, B., Sarimanah, E., Dewi, F. I., & Sabri, T. (2019). A Review of Students' Common Misconceptions in Science and Their Diagnostic Assessment Tools. *Jurnal Pendidikan IPA Indonesia*, *8*(2). https://doi.org/10.15294/jpii.v8i2.18649

Sosa-Moguel, L., & Aparicio-Landa, E. (2021). Secondary school mathematics teachers' perceptions about inductive reasoning and their interpretation in teaching. *Journal on Mathematics Education*, *12*(2), 239–256. https://doi.org/10.22342/JME.12.2.12863.239-256

Steedle, J. T., & Shavelson, R. J. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *46*(6), 699–715.

Stefanidou, C. G., Tsalapati, K. D., Ferentinou, A. M., & Skordoulis, C. D. (2019). Conceptual Difficulties Pre-Service Primary Teachers Have with Static Electricity. *Journal of Baltic Science Education*, *18*(2), 300.

Stephens, R. G., Dunn, J. C., Hayes, B. K., & Kalish, M. L. (2020). A test of two processes: The effect of training on deductive and inductive reasoning. *Cognition*, *199*(February), 104223. https://doi.org/10.1016/j.cognition.2020.104223

Sternberg, R. J., Sternberg, K., & Mio, J. (2012). *Cognitive psychology*. Cengage Learning Press.

Strobel, A., Behnke, A., Gärtner, A., & Strobel, A. (2019). The interplay of intelligence and need for cognition in predicting school grades: A retrospective study. *Personality and Individual Differences*, *144*(March), 147–152. https://doi.org/10.1016/j.paid.2019.02.041

Subali, B., Kumaidi, Aminah, N. S., & Sumintono, B. (2019). Student achievement based on the use of scientific method in the natural science subject in elementary school. *Jurnal Pendidikan IPA Indonesia*, *8*(1), 39–51. https://doi.org/10.15294/jpii.v8i1.16010

Sukarelawan, M. I., Jumadi, J., Kuswanto, H., Soeharto, S., & Hikmah, F. N. (2021). Rasch Analysis to Evaluate the Psychometric Properties of Junior Metacognitive Awareness Inventory in the Indonesian Context. *Jurnal Pendidikan IPA Indonesia*, *10*(4), Article 4. https://doi.org/10.15294/jpii.v10i4.27114

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi Model Rasch untuk Penelitian Ilmu-ilmu Sosial*. Trim Komunikata Publishing House.

Tabachnick, B. G., & Fidell, L. S. (2019). *Using Multivariate Statistics*. Pearson: Boston, MA, USA.

Taber, K. S. (2018). The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. *Research in Science Education*, *48*(6), 1273–1296. https://doi.org/10.1007/s11165-016-9602-2

Taskinen, P. H., Dietrich, J., & Kracke, B. (2015). The role of parental values and child-specific expectations in the science motivation and achievement of adolescent girls and boys. *International Journal of Gender, Science and Technology*, *8*(1), 103–123.

Taslidere, E. (2016). Development and use of a three-tier diagnostic test to assess high school students' misconceptions about the photoelectric effect. *Research in Science & Technological Education*, *34*(2), 164–186. https://doi.org/10.1080/02635143.2015.1124409

Tavakol, M., & Dennick, R. (2013). Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE Guide No. 72. *Medical Teacher, 35*(1), e838–e848. https://doi.org/10.3109/0142159X.2012.737488

Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J., & Janssen, R. (2017). Measuring Critical Thinking in Physics: Development and Validation of a Critical Thinking Test in Electricity and Magnetism. *International Journal of Science and Mathematics Education*, *15*(4), 663–682. https://doi.org/10.1007/s10763-016-9723-0

Topalsan, A. K., & Bayram, H. (2019). Identifying Prospective Primary School Teachers' Ontologically Categorized Misconceptions on the Topic of" Force and Motion". *Journal of Turkish Science Education, 16*(1), 85–109.

Treagust, D. F. (1986). Evaluating students' misconceptions by means of diagnostic multiple choice items. *Research in Science Education*, *16*(1), 199–207. https://doi.org/10.1007/BF02356835

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159–169. https://doi.org/10.1080/0950069880100204

Treagust, D. F., & Duit, R. (2008). Conceptual change: A discussion of theoretical, methodological and practical challenges for science education. *Cultural Studies of Science Education*, *3*(2), 297–328. https://doi.org/10.1007/s11422-008-9090-4

Treagust, D. F., Mthembu, Z., & Chandrasegaran, A. L. (2014). Evaluation of the Predict-Observe-Explain Instructional Strategy to Enhance Students' Understanding of Redox Reactions. In *Learning with Understanding in the Chemistry Classroom* (pp. 265–286). Springer Netherlands. https://doi.org/10.1007/978-94-007-4366-3_14

Tsui, C., & Treagust, D. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *International Journal of Science Education*, *32*(8), 1073–1098.

Tümay, H. (2016). Reconsidering learning difficulties and misconceptions in chemistry: Emergence in chemistry and its implications for chemical education. *Chemistry Education Research and Practice, 17*(2), 229–245. https://doi.org/10.1039/c6rp00008h

Van Vo, D., & Csapó, B. (2020). Development of inductive reasoning in students across school grade levels. *Thinking Skills and Creativity*, *37*(March), 1–16. https://doi.org/10.1016/j.tsc.2020.100699

Van Vo, D., & Csapó, B. (2021). Exploring students' science motivation across grade levels and the role of inductive reasoning in science motivation. *European Journal of Psychology of Education*. https://doi.org/10.1007/s10212-021-00568-8

Van Vo, D., & Csapó, B. (2022). Exploring students' science motivation across grade levels and the role of inductive reasoning in science motivation. *European Journal of Psychology of Education*, *37*(3), 807–829.

Van Vo, D., & Csapó, B. (2023). Exploring Inductive Reasoning, Scientific Reasoning and Science Motivation, and Their Role in Predicting STEM Achievement Across Grade Levels. *International Journal of Science and Mathematics Education*, 1–24.

Venville, G., & Oliver, M. (2015). The impact of a cognitive acceleration programme in science on students in an academically selective high school. *Thinking Skills and Creativity*, *15*, 48–60. https://doi.org/10.1016/j.tsc.2014.11.004

Wandersee, J. H., Mintzes, J. J., & Novak, J. D. (1994). Research on alternative conceptions in science. *Handbook of Research on Science Teaching and Learning*, *177*, 210.

Waschl, N., & Burns, N. R. (2020). Sex differences in inductive reasoning: A research synthesis using meta-analytic techniques. *Personality and Individual Differences*, *164*(February), 109959. https://doi.org/10.1016/j.paid.2020.109959

Wickham, H. (2016). Data analysis. In *Ggplot2* (pp. 189–201). Springer.

Wu, H., & Molnár, G. (2018). Interactive problem solving: Assessment and relations to combinatorial and inductive reasoning. *Journal of Psychological and Educational Research*, *26*(1), 90–105.

Wyse, A. E., & Mapuranga, R. (2009). Differential Item Functioning Analysis Using Rasch Item Information Functions. *International Journal of Testing*, *9*(4), 333–357. https://doi.org/10.1080/15305050903352040

Yilmaz, S., Eryilmaz, A., & Geban, O. (2006). Assessing the Impact of Bridging Analogies in Mechanics. *School Science and Mathematics*, *106*(6), 220–230. https://doi.org/10.1111/j.1949-8594.2006.tb17911.x

Yin, R. K. (2018). *Case study research and applications: Design and methods*. Sage Books.

Zhu, X., & Neupert, S. D. (2021). Dynamic awareness of age-related losses predict concurrent and subsequent changes in daily inductive reasoning performance. *British Journal of Developmental Psychology*, *39*(2), 282–298. https://doi.org/10.1111/bjdp.12344

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The international state of research on measurement of competency in higher education. *Studies in Higher Education*, *40*(3), 393–411. https://doi.org/10.1080/03075079.2015.1004241

Zook, K. B., & Maier, J. M. (1994). Systematic analysis of variables that contribute to the formation of analogical misconceptions. *Journal of Educational Psychology*, *86*(4), 589–600. https://doi.org/10.1037/0022-0663.86.4.589

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*(1), 1–28.

APPENDICES

Appendix 1: letter of consent

Research title: Analyzing students' misconception on science education and factors influencing them in the context of Indonesian learner.

Dear Participant/students

You are participating in the research conducted by Soeharto Soeharto from the University of Szeged (SZTE). This research has been approved by Institutional Review Board of Doctoral School of Education, SZTE. Read each question carefully and answer as accurately as possible. We inform you that your response will be processed anonymously with other participants. We keep all information confidential during research. Research records and all information will be encoded and secured using password protected files. Data analysis will be carried out without entering the respondent's identity. You will never be identified in this research project or in other presentations or publications. The information you provide will be coded only by numbers. Research results will be submitted for publication in scientific journals or presented at scientific conferences. Participation in the research is anonymous and voluntary.

You may ask if you do not understand or are unsure of how answer it to researcher or send us email to soeharto.soeharto@edu.u-szeged.hu

Thank you for considering participation in the research.

Online form

You can give your consent by ticking the boxes below. Thank you for agreeing to participate in the research.

□	I have read the information about this research questionnaires, and any questions that I wanted to ask have been answered clearly
□	I agree to participate in this research, and I understand that I can withdraw my consent at any time, before the end of my participation

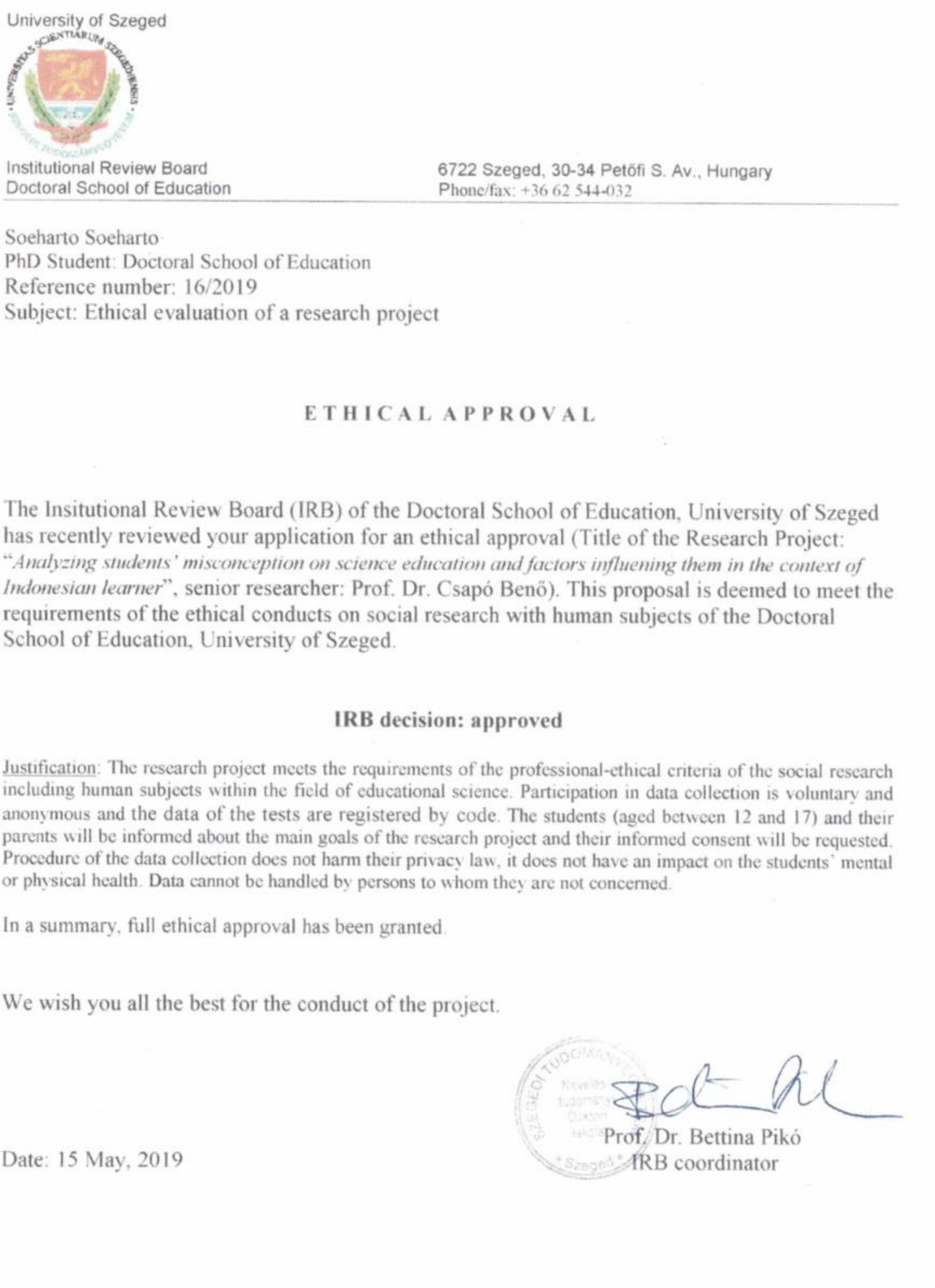
Paper form

I …..(Name)..... agree to participate in the study.

Signature:…………………………….. Date:………………………………………

Appendix 2

Appendix 2: ethical approval from the institutional review board of the university of szeged



University of Szeged

Institutional Review Board
Doctoral School of Education

6722 Szeged, 30-34 Petőfi S. Av., Hungary
Phone/fax: +36 62 544-032

Soeharto Soeharto
PhD Student: Doctoral School of Education
Reference number: 16/2019
Subject: Ethical evaluation of a research project

ETHICAL APPROVAL

The Insitutional Review Board (IRB) of the Doctoral School of Education, University of Szeged has recently reviewed your application for an ethical approval (Title of the Research Project: "*Analyzing students' misconception on science education and factors influening them in the context of Indonesian learner*", senior researcher: Prof. Dr. Csapó Benő). This proposal is deemed to meet the requirements of the ethical conducts on social research with human subjects of the Doctoral School of Education, University of Szeged.

IRB decision: approved

Justification: The research project meets the requirements of the professional-ethical criteria of the social research including human subjects within the field of educational science. Participation in data collection is voluntary and anonymous and the data of the tests are registered by code. The students (aged between 12 and 17) and their parents will be informed about the main goals of the research project and their informed consent will be requested. Procedure of the data collection does not harm their privacy law, it does not have an impact on the students' mental or physical health. Data cannot be handled by persons to whom they are not concerned.

In a summary, full ethical approval has been granted.

We wish you all the best for the conduct of the project.

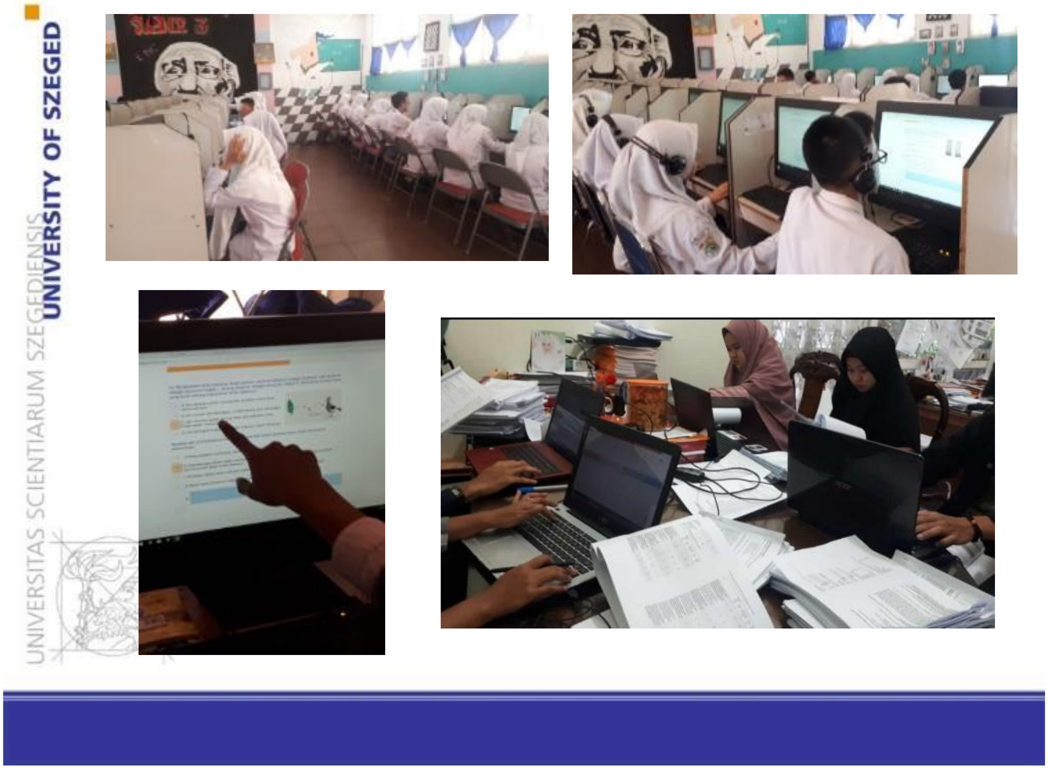
SZEGEDI TUDOMÁNYEGYETEM
Neveléstudományi
Doktori
Iskola
Szeged

Prof. Dr. Bettina Pikó
IRB coordinator

Date: 15 May, 2019

Appendix 3: data collection photos

Documentation



Appendix 4: the two-tier multiple-choice test

Tes Pilihan Ganda Dua Tingkat Pada Mata Pelajaran Sains

FISIKA

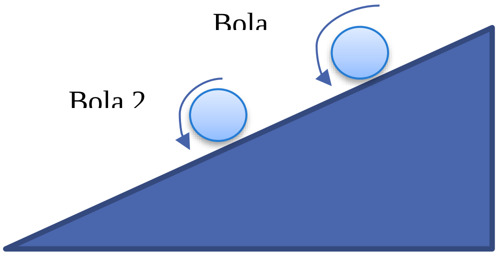
1. Dua orang Pelari berpartisipasi dalam sebuah lomba. Pelari 1 memiliki energi kinetik lebih besar dari Pelari 2. Apakah Pelari 1 memiliki berat lebih dari, kurang dari, atau sama dengan Pelari 2?

   a) Pelari 1 beratnya sama dengan Pelari 2.
   b) Pelari 1 lebih berat dari Pelari 2.
   c) Pelari 1 memiliki berat kurang dari Pelari 2.
   d) Satu-satunya cara untuk mengetahui Pelari mana yang lebih berat adalah dengan mengetahui seberapa kecepatan setiap Pelari.

   Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

   a) jika dua benda bergerak dengan kecepatan yang sama, yang lebih berat akan memiliki lebih sedikit energi kinetik.
   b) jika dua benda bergerak dengan kecepatan yang sama, yang lebih berat akan memiliki lebih banyak energi kinetik.
   c) benda yang lebih berat terbebani, sehingga tidak memiliki banyak energi kinetik.
   d) benda yang lebih ringan tidak terbebani, sehingga memiliki banyak energi kinetik.
   e) ....

2. Dua bola dengan yang identik bergulir di bidang miring. Bola 2 lebih cepat dari Bola 1.



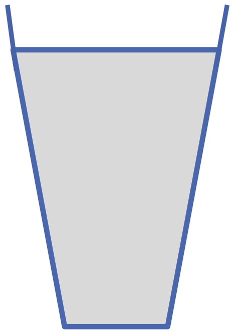
   Bola mana yang memiliki energi kinetik lebih besar?
   a) Bola 2 memiliki energi kinetik yang lebih besar.
   b) Bola 1 memiliki energi kinetik yang lebih besar.
   c) Bola 1 dan Bola 2 memiliki jumlah energi kinetik yang sama.
   d) Kedua bola tidak memiliki energi kinetik.

   Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

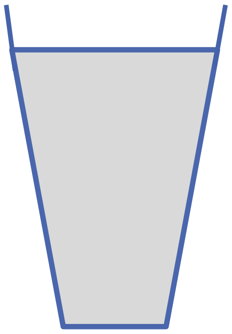
   a) Energi kinetik bola tidak tergantung pada kecepatan
   b) Energi kinetik bola tergantung pada kecepatan, massa dan posisi

c) Energi kinetik bola tergantung pada kecepatan
d) Energi kinetik bola tergantung pada kecepatan dan posisi karena kedua bola identik
e) ....

3. Seorang siswa memiliki dua gelas susu yang identik. Awalnya suhu susu di gelas itu sama. Setelah disimpan di tempat yang berbeda untuk sementara waktu, suhu susu berubah. Suhu susu dalam Gelas 1 adalah 30°C, dan suhu susu dalam Gelas 2 adalah 70°C



Gelas 1 – Susu, 30°C



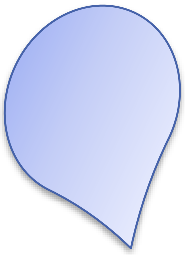
Gelas 2 –Susu, 70°C

Manakah gelas susu yang memiliki lebih banyak energi termal?

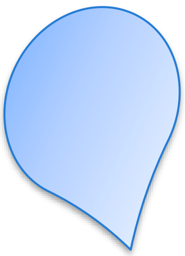
a) Susu pada suhu 30°C memiliki lebih banyak energi termal.
b) Susu pada 30°C dan susu pada 70°C memiliki jumlah energi termal yang sama.
c) Susu pada suhu 70°C memiliki lebih banyak energi termal.
d) Susu pada 30°C dan susu pada 70°C tidak memiliki energi termal.

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?
a) Jumlah energi termal suatu benda berkurang ketika suhu benda meningkat.
b) Suatu zat memiliki energi termal yang sama terlepas dari suhunya atau keadaannya.
c) Energi termal suatu benda tidak terkait dengan suhu benda tersebut.
d) Jumlah energi panas suatu benda meningkat ketika suhu benda meningkat.
e) ....

4. Seorang anak memiliki dua balon berisi gas Helium. Balon 1 dan Balon 2 memiliki jumlah atom helium yang sama.



Balon 1



Balon 2

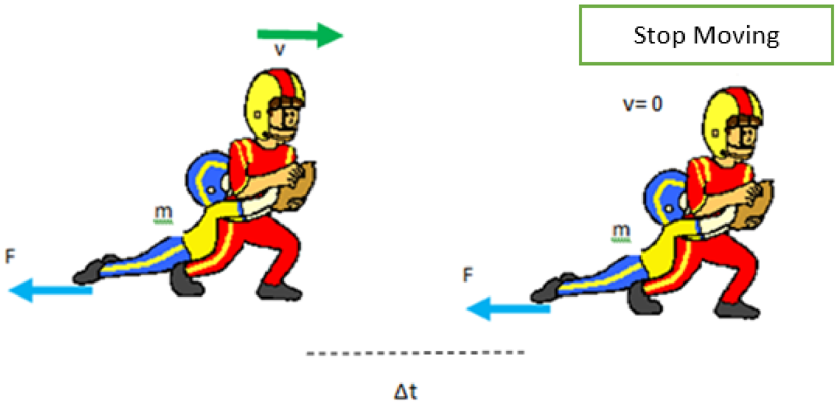
Jika energi termal helium di Balon 1 ditingkatkan sehingga Balon 1 memiliki lebih banyak energi termal daripada helium di Balon 2. Atom helium mana yang akan bergerak lebih cepat dari rata-rata?

a) Atom helium di Balon 2 akan bergerak lebih cepat dari rata-rata.
b) Atom helium di Balon 1 akan bergerak lebih cepat dari rata-rata
c) Atom helium di Balon 1 akan bergerak dengan kecepatan rata-rata yang sama dengan atom helium di Balon 2.
d) Satu-satunya cara untuk mengetahui atom helium mana yang akan bergerak lebih cepat rata-rata adalah dengan mengetahui suhu helium di setiap balon.

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Energi panas tidak terkait dengan kecepatan molekul yang membentuk suatu objek.
b) Energi panas berhubungan dengan kecepatan molekul yang membentuk suatu objek
c) Energi panas suatu benda terkait dengan jenis gas benda tersebut.
d) Jumlah energi panas suatu benda mengalami penurunan sejalan dengan meningkatnya kecepatan benda tersebut.
e) ....

5. Pemain American football biru memegang pemain American football merah yang bergerak dengan kecepatan (v), sambil membaringkan tubuhnya ke tanah. Pemain amerikan football biru mengunakan gaya gesek tubuh (F) terhadap tanah dan massa tubuh (m) sehingga dalam selang waktu (Δt), pemain Amerikan football merah berhenti bergerak (perhatikan gambar dengan seksama). Dari peristiwa ini, apakah hubungan antara impuls dan momentum?



a) Impuls sama dengan momentum
b) Impuls sama dengan perubahan momentum
c) impuls mengurangi momentum
d) impuls meningkatkan momentum

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Dari gambar, dapat dilihat bahwa kecepatan pemain sepak bola telah menurun sehingga momentumnya juga akan berkurang.
b) Dari gambar, dapat dilihat bahwa kecepatan pemain sepak bola telah menurun sehingga momentumnya juga akan meningkat.
c) Dari gambar, momentum berubah pada interval singkat karena gesekan tubuh pemain sepak bola dengan tanah.
d) Dari gambar, momentumnya sama dengan impuls karena dalam kasus yang sama pada sistem transfer energi.
e) ....

6. Dua kelereng memiliki massa yang sama, Kelerang A dan Kelerang B berada pada permukaan datar yang licin (tanpa gesekan). Kelerang A dan Kelerang B bergerak saling mendekat satu sama lain dengan kecepatan tertentu. Setelah bertumbukan kedua kelereng terpisah dan bergerak ke arah yang berlawanan. pernyataan mana yang benar tentang momentum total benda sebelum dan sesudah tabrakan?

a) Momentum total sebelum tumbukan sama dengan Momentum total setelah tumbukan
b) Momentum total sebelum tumbukan lebih besar dari momentum total setelah tumbukan
c) Momentum total sebelum tumbukan lebih kecil dari momentum total setelah tumbukan
d) Satu-satunya cara untuk mengetahui momentum total sebelum dan sesudah tumbukan adalah dengan mengetahui kecepatan dan massa benda.

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Setelah tumbukan, energi hilang, sehingga momentum total setelah tumbukan akan berkurang.
b) Setelah tumbukan, energi naik, sehingga Momentum total setelah tumbukan akan meningkat.
c) Momentum total sebelum dan sesudah tumbukan adalah sama berdasarkan massa benda.
d) Momentum total sebelum dan sesudah tabrakan adalah sama berdasarkan pada hukum konservasi momentum.
e) ....

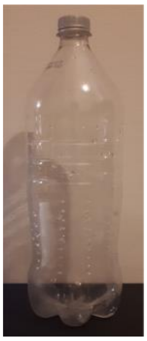
7. Seorang ibu mencuci baju dan kemudian menggantung baju yang basah di jemuran di bawah sinar matahari. beberapa jam kemudian bajunya kering. Apa yang terjadi pada molekul air di baju itu?

a) Molekul air menjadi bagian dari baju itu.
b) Molekul air menghilang karena sinar matahari.
c) Molekul air diubah menjadi atom hidrogen dan oksigen.
d) Molekul air bergerak lebih cepat dan menjadi bagian dari udara

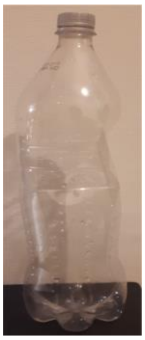
Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Ketika air menguap dari baju, molekul air diserap oleh baju.
b) Molekul air dihancurkan selama proses penguapan.
c) Molekul air berubah menjadi atom individu ketika proses penguapan.
d) Molekul air bergerak lebih cepat dan berubah menjadi gas sebagai bagian dari udara.
e) ....

8. Seorang siswa minum air dari botol plastik hingga habis. Kemudian, siswa menutup botol plastik dengan kuat dan memasukkannya ke dalam kulkas. Satu jam kemudian botol itu terlihat penyok. Apakah yang menyebabkan botol penyok setelah didinginkan di lemari es?



Botol sebelum didinginkan



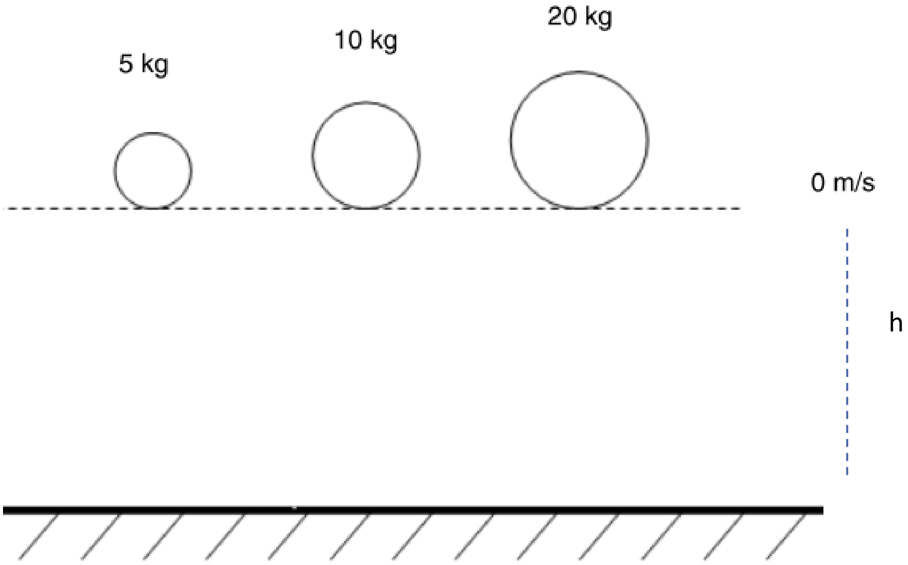
Botol setelah didinginkan

a) Semua molekul udara mencoba keluar dari botol.
b) Udara panas dalam botol dihancurkan oleh suhu dingin.
c) Molekul-molekul udara di dalam botol menghilang karena perbedaan tekanan
d) Molekul-molekul udara di dalam botol semakin berdekatan.

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Semua molekul akan hancur karena suhu dingin.
b) Semua molekul dalam kasus itu berubah menjadi energi panas.
c) Semua molekul dapat dihancurkan oleh perbedaan tekanan.
d) Penurunan suhu biasanya mengurangi jarak antara atom atau molekul.
e) ....

9. Tiga bola pejal dengan massa berbeda, 5 kg, 10 kg, dan 20 kg, jatuh dari ketinggian tertentu (h) secara bersamaan hingga menyentuh tanah. Berdasarkan situasi ini, jika efek hambatan udara dihilangkan, manakah dari pernyataan ini yang benar?

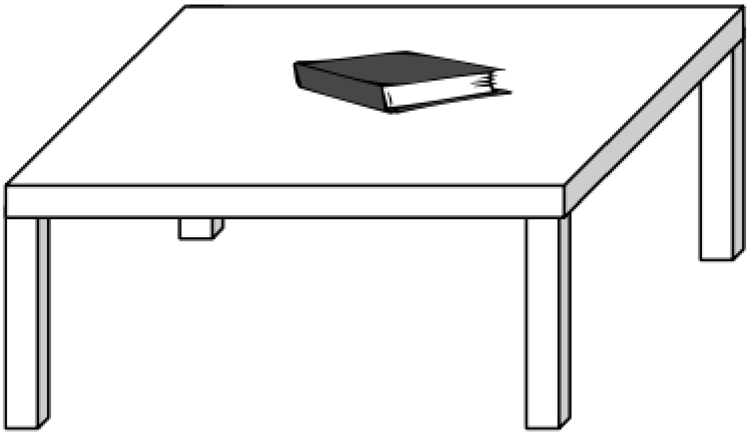


a) Bola 5 kg akan menyentuh tanah terlebih dahulu
b) Bola 10 kg akan menyentuh tanah terlebih dahulu
c) Bola 20 kg akan menyentuh tanah terlebih dahulu
d) Semua bola akan menyentuh tanah bersamaan

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Bola berat jatuh dengan kecepatan lebih besar dari bola ringan.
b) Bola ringan jatuh dengan kecepatan lebih besar dari bola berat.
c) Bola jatuh dengan kecepatan yang sama tidak tergantung dari massanya
d) Bola berat jatuh dengan kecepatan lebih besar dan menyentuh tanah terlebih dahulu karena tarikan gravitasi
e) ....

10. Buku dengan berat 10 N ditempatkan di atas meja seperti yang ditunjukkan pada gambar. Buku itu dalam kondisi diam. pernyataan manakah yang benar?



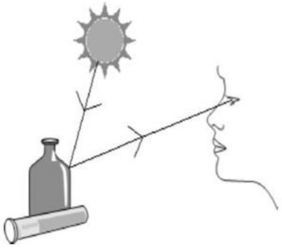
a) Buku pada kondisi diam tidak memiliki gaya yang bekerja.
b) Buku itu hanya memiliki gaya berat 10 N dan gaya reaksi 10 N
c) Buku itu hanya memiliki gaya kontak 10 N
d) Buku memiliki gaya berat, gaya kontak, dan gaya reaksi dengan jumlah yang sama, masing-masing gaya memiliki 10 N.

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

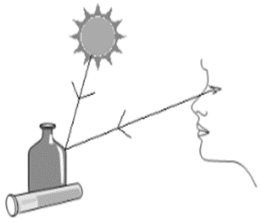
a) Ketika resultan gaya pada buku adalah nol, tidak ada gaya yang bekerja pada buku.
b) Semua gaya yang bekerja pada buku memiliki jumlah yang sama, tetapi gaya yang bekerja pada buku tersebut tidak nol karena buku tertahan di meja.
c) Gaya yang dihasilkan dalam buku sama dengan jumlah semua gaya yang bekerja pada sistem buku.
d) Ketika buku dalam kondisi diam, semua gaya saling meniadakan
e) ....

11. Manakah dari gambar berikut yang dengan benar menunjukkan proses agar sebuah objek dapat dilihat oleh mata manusia?

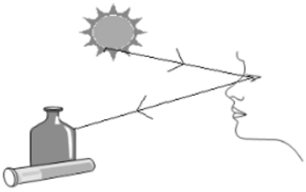
a) Gambar A



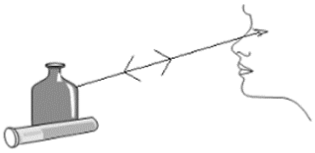
b) Gambar B



c) Gambar C



d) Gambar D

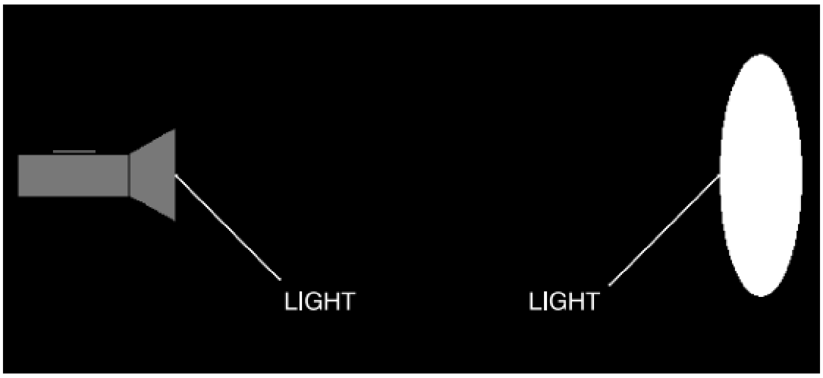


Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Kita dapat melihat suatu objek karena cahaya datang ke objek dan datang ke mata manusia, sehingga menyebabkan rangsangan dalam bentuk sinyal kimia dan listrik ke otak.
b) Kita dapat melihat suatu objek karena cahaya datang ke mata manusia dan memantulkannya ke objek, sehingga menimbulkan rangsangan berupa sinyal kimia dan listrik ke otak.

c) Kita dapat melihat suatu objek karena mata manusia mampu menerima sinar cahaya yang berasal dari benda, sehingga menimbulkan rangsangan berupa sinyal kimia dan listrik ke otak.
d) Kita dapat melihat objek karena cahaya ada antara objek dan mata manusia.
e) ....

12. Anda menggunakan senter di ruangan untuk menerangi permukaan gelap dinding seperti yang ditunjukkan pada gambar. Berdasarkan sifat cahaya, manakah dari pernyataan berikut yang benar?



a) Cahaya hanya pada area permukaan yang terang.
b) Cahaya ada pada bohlam senter dan pada area permukaan yang terang.
c) Cahaya hanya pada bola lampu senter yang merupakan sumber cahaya.
d) Cahaya ada pada bola lampu senter hingga ke area permukaan yang terang.

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Cahaya hanya dapat ditemukan di area yang tdapat dilihat.
b) Cahaya adalah paket.paket sinar yang ada di area yang terlihat..
c) Kita dapat melihat cahaya di area yang terang karena cahaya dalam bentuk sinar-sinar bergerak melintasi ruangan dengan kecepatan sangat tinggi.
d) Cahaya tidak ada pada malam hari, sehingga cahaya hanya ada pada sumber cahaya yang merupakan area terang.
e) ....

BIOLOGY

13. Manakah sel yang memiliki fungsi dasar sel-sel pada proses pertumbuhan?

a) Sel-sel hewan, tetapi bukan sel-sel tumbuhan
b) Sel tanaman, tetapi bukan sel hewan
c) Sel hewan dan sel tumbuhan
d) Bukan sel tumbuhan maupun sel hewan

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Sel-sel hewan tidak berperan dalam fungsi proses pertumbuhan
b) Sel-sel tumbuhan tidak berperan dalam fungsi proses pertumbuhan
c) Sel-sel tumbuhan tidak menghasilkan molekul yang berguna untuk pertumbuhan
d) Sel-sel tumbuhan dan hewan memiliki fungsi penting dalam proses pertumbuhan
e) ....

14. Manakah dari pernyataan berikut ini yang benar tentang sel-sel otot pada hewan?

a) Sel-sel otot mendapatkan energi dari makanan yang digunakan untuk membuat molekul untuk pertumbuhan.
b) Sel-sel otot mendapatkan energi dari makanan namun tidak digunakan untuk membuat molekul untuk pertumbuhan..
c) Sel-sel otot membuat molekul untuk pertumbuhan dari protein di bagian tubuhnya.
d) Muscle cells do not produce molecules for growth, but they are not used to get energy from food.

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Sel-sel hewan tidak mengekstraksi energi dari makanan.
b) Sel-sel hewan tidak menjalankan fungsi kehidupan yang penting bagi diri mereka sendiri.
c) Sel-sel hewan tidak membuat molekul untuk pertumbuhannya sendiri.
d) Sel-sel hewan membuat molekul untuk pertumbuhannya sendiri.
e) ....

15. Seorang siswa menghirup udara di lapangan. Siswa itu berulang kali menghirup dan menghembuskan udara dalam beberapa menit. Manakah dari pernyataan berikut ini yang benar?

a) Siswa menghirup oksigen dan menghembuskan karbon dioksida.
b) Siswa menghirup karbon dioksida dan menghembuskan oksigen.
c) Siswa menghirup oksigen dan menghembuskan karbon dioksida dan dihidrogen monoksida.
d) Siswa menghirup berbagai komponen udara dalam bentuk oksigen, karbon dioksida dan lainnya, dan mengeluarkan berbagai komponen udara dalam bentuk karbon dioksida, oksigen, dan lainnya.

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Udara yang dihembuskan banyak mengandung karbon dioksida dan sangat sedikit oksigen.
b) Manusia memberi karbon dioksida ke tumbuhan dan tumbuhan memberi oksigen kepada manusia.

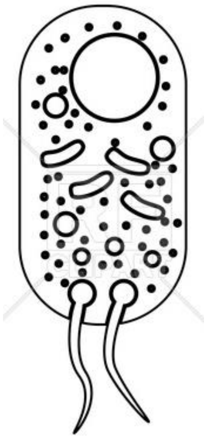
c) Udara yang dihirup biasanya akan mengandung uap air, dan jumlahnya tergantung pada kelembaban lingkungan.
d) Dalam proses bernapas pada manusia, udara memiliki beberapa komponen seperti oksigen, karbon dioksida, nitrogen dan lainnya.
e) ....

16. Paru-paru adalah organ penting dalam sistem pernapasan kita. Manakah dari pernyataan berikut ini yang benar mengenai pergerakan oksigen antara paru-paru dan sel-sel tubuh manusia?

a) Sebagian besar molekul oksigen bergerak dari sel-sel tubuh ke paru-paru dengan melalui sistem peredaran darah dan kemudian meninggalkan sistem peredaran darah melalui pembuluh kapiler di paru-paru.
b) Sebagian besar molekul oksigen bergerak dari paru-paru ke jantung melalui saluran pernapasan khusus dimana molekul oksigen bercampur dengan darah. Darah kemudian bergerak ke sel-sel tubuh melalui sistem peredaran darah.
c) Molekul oksigen masuk dan meninggalkan tubuh melalui paru-paru, tetapi mereka tidak bergerak antara paru-paru dan darah.
d) Sebagian besar molekul oksigen bergerak dari paru-paru ke sel-sel tubuh dengan memasukkan pembuluh darah kapiler secara mikroskopis dan kemudian pindah ke sel melalui sistem peredaran darah.

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?
a) Sistem pernapasan dan sistem sirkulasi tidak terhubung.
b) Jantung adalah tempat pencampuran udara dan darah.
c) Semua molekul udara bergerak dari paru-paru ke darah dengan meelalui pembuluh kapiler.
d) Sistem pernapasan dan sistem sirkulasi terhubung, tetapi sebagian besar molekul oksigen tidak masuk ke kapiler di paru-paru
e) ....

17. Gambar di bawah ini adalah salah satu jenis bakteri. Manakah dari pernyataan berikut ini yang benar tentang sistem kehidupan bakteri?

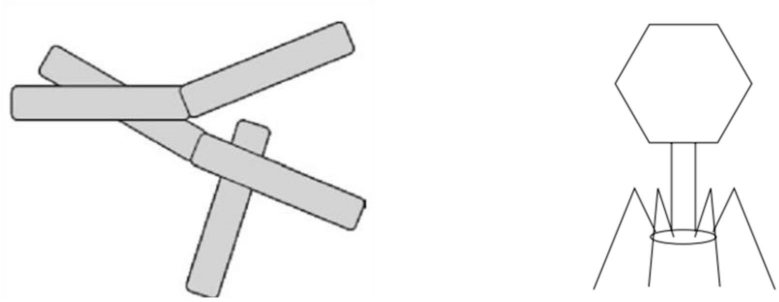


a) Bakteri menggunakan paru-paru untuk bernafas
b) Bakteri menggunakan usus sebagai sistem pencernaan
c) Bakteri menggunakan mulutnya untuk mengikat oksigen
d) Bakteri menggunakan sistem difusi melalui sel membran untuk mendapatkan nutrisi

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Bakteri mempunyai organ vital untuk bertahan hidup
b) Bakteri terdiri dari satu sel tunggal
c) Sistem pernapasan dan pencernaan sama dengan hewan lainnya
d) Bakteri adalah organisme multiseluler
e) ....

18. Gambar di bawah ini adalah gambar bakteri E.coli dan virus Phage. Manakah pernyataan yang tepat terkait dengan peran antibiotik pada bakteri dan virus?



Bakteri E. coli

Virus Phage

a) Antibiotik membunuh bakteri dan virus dengan mencegah reaksi kimia internal.
b) Antibiotik membunuh virus dengan mencegah reaksi kimia internal.
c) Antibiotik membunuh bakteri dengan mencegah reaksi kimia internal
d) Virus dapat disembuhkan dengan antibiotik melalui interaksi kimia

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Virus terdiri dari asam nukleat dan protein melalui reaksi kimia.
b) Antibiotik dapat mengganggu keseimbangan proses dalam sel.
c) Virus memiliki reaksi kimia internal, sehingga dapat dipengaruhi oleh antibiotik.
d) Bakteri adalah organisme hidup, tetapi Virus bukanlah organisme hidup.
e) ....

19. Manakah dari pernyataan berikut ini yang benar terkait dengan sistem pencernaan?

    a) Pencernaan diperlukan untuk memecah protein dan karbohidrat kompleks menjadi molekul yang cukup kecil agar bisa masuk ke sel-sel tubuh.
    b) Pencernaan diperlukan untuk memecah protein menjadi molekul yang cukup kecil untuk masuk ke sel-sel tubuh, tetapi tidak diperlukan untuk memecah karbohidrat kompleks karena mereka sudah cukup kecil untuk masuk ke sel-sel tubuh
    c) Pencernaan diperlukan untuk memecah karbohidrat kompleks menjadi molekul yang cukup kecil untuk masuk ke sel-sel tubuh, tetapi tidak diperlukan untuk memecah protein karena mereka sudah cukup kecil untuk masuk ke sel-sel tubuh.
    d) Pencernaan tidak diperlukan untuk memecah baik protein atau karbohidrat kompleks karena molekul-molekul ini sudah cukup kecil untuk masuk ke sel-sel tubuh.

    Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

    a) Karbohidrat kompleks tidak harus dipecah menjadi molekul yang lebih kecil sebelum mereka dapat masuk ke dalam sel-sel tubuh.
    b) Karbohidrat kompleks dan protein harus dipecah menjadi molekul yang lebih kecil sebelum mereka dapat memasuki sel-sel tubuh.
    c) Protein tidak harus dipecah menjadi molekul yang lebih kecil sebelum mereka dapat memasuki sel-sel tubuh.
    d) Molekul dari makanan didistribusikan melalui tabung khusus dalam tubuh, sehingga karbohidrat kompleks dapat memasuki sel-sel tubuh
    e) ....

20. Bagaimana molekul makanan dan molekul oksigen mencapai sel-sel dalam tubuh?

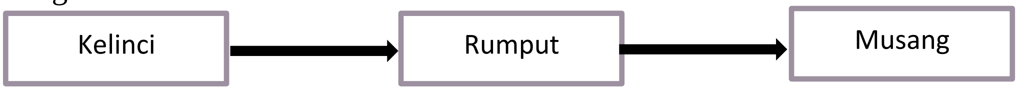
    a) Molekul dari makanan dan molekul oksigen dibawa oleh serangkaian tabung perantara yang menghubungkan mulut dan hidung ke seluruh bagian tubuh.
    b) Molekul dari makanan dan molekul oksigen dibawa oleh serangkaian tabung perantara yang menghubungkan lambung dan paru-paru ke seluruh tubuh.
    c) Molekul dari makanan dan molekul oksigen dibawa oleh jaringan arteri, vena, dan pembuluh darah kecil secara mikroskopis melalui pembuluh ke seluruh tubuh.
    d) Molekul dari makanan dan molekul oksigen bergerak langsung dari mulut dan hidung ke seluruh tubuh tanpa melalui perantara apapun.

    Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

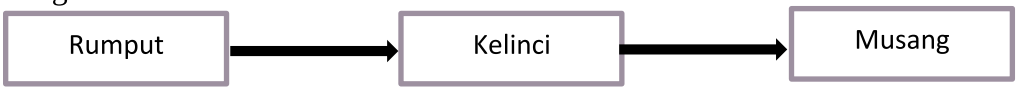
a) Molekul dari makanan bergerak langsung dari mulut ke seluruh tubuh tanpa melalui tabung apa pun.
b) Makanan secara bebas masuk ke dalam tubuh (tidak ada kaitannya dengan struktur pencernaan).
c) Molekul dari makanan didistribusikan melalui tabung khusus, bukan melalui sistem sirkulasi ke seluruh tubuh.
d) Sistem peredaran darah adalah jalan yang digunakan untuk membawa molekul makanan dan oksigen ke seluruh tubuh.
e) ....

21. Manakah dari diagram rantai makanan berikut ini yang benar?

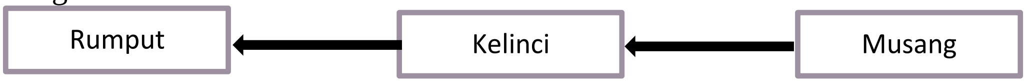
a) Diagram A



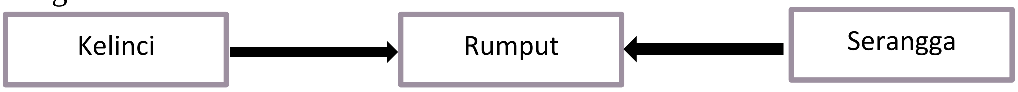
b) Diagram B



c) Diagram C



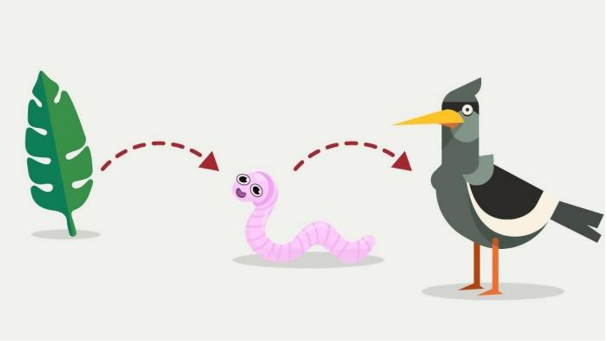
d) Diagram D



Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Tanda panah dalam rantai makanan berarti 'memakan'.
b) Rantai makanan adalah transfer energi makanan dari sumber daya berupa tanaman melalui serangkaian organism.
c) Musang adalah predator dalam rantai makanan
d) Hewan yang kuat harus berada di puncak rantai makanan
e) ....

22. Berdasarkan rantai makanan dalam gambar, tanaman berperan sebagai produsen, ulat berperan sebagai konsumen tingkat I, burung berperan sebagai konsumen tingkat II. Pernyataan berikut mana yang benar tentang mekanisme rantai makanan?



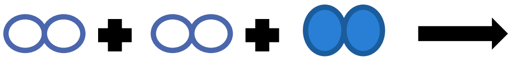
a) jika tanaman punah, burung akan bertahan hidup dalam rantai makanan.
b) jika jumlah ulat meningkat, jumlah burung akan meningkat.
c) jika tanaman punah, burung tidak akan bertahan hidup dalam rantai makanan.
d) jika burung-burung punah, rantai makanan masih seimbang.

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

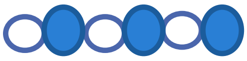
a) Tanpa predator (karnivora) rantai makanan akan tetap seimbang
b) Keseimbangan dalam rantai makanan sangat penting dalam kaitannya dengan jumlah produsen dan konsumen dalam rantai makanan.
c) Produsen dalam rantai makanan tidak mempengaruhi jumlah konsumen di tingkat puncak.
d) Tanpa herbivora rantai makanan akan tetap seimbang.
e) ....

KIMIA

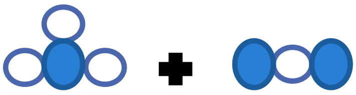
23. Gambar berikut menunjukkan molekul-molekul sebelum reaksi kimia terjadi. atom diwakili oleh lingkaran dan molekul diwakili oleh lingkaran yang terhubung satu sama lain. Gambar apa yang menunjukkan molekul yang dihasilkan dari reaksi kimia?



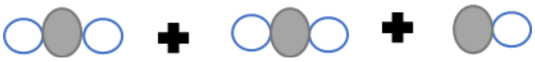
e) Gambar A



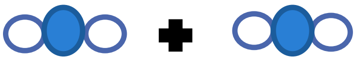
f) Gambar B



g) Gambar C



h) Gambar D

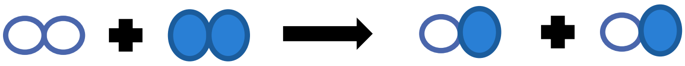


Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

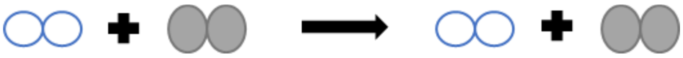
a) Ada 6 atom sebelum reaksi dan 6 atom setelah reaksi.
b) Ada 4 atom putih dan 2 atom abu-abu sebelum reaksi, dan 4 atom putih dan 2 atom abu-abu setelah reaksi.
c) Ada 2 jenis molekul sebelum reaksi dan 2 jenis molekul setelah reaksi.
d) Ada 3 molekul sebelum reaksi dan 3 molekul setelah reaksi.
e) ....

24. Manakah dari gambar berikut yang dapat dengan tepat menggambarkan reaksi kimia?
Atom diwakili oleh lingkaran, dan molekul diwakili oleh lingkaran yang terhubung satu sama lain. Lingkaran berwarna yang berbeda mewakili berbagai jenis atom.

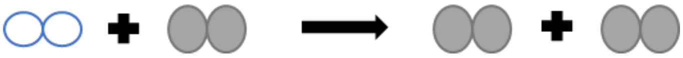
a) Gambar A



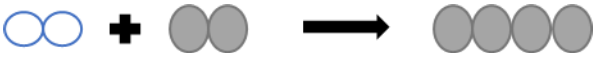
b) Gambar B



c) Gambar C



d) Gambar D

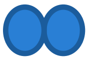


Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

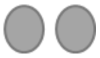
a) Atom-atom reaktan dari suatu reaksi kimia diubah menjadi atom-atom lain.
b) Setelah reaksi kimia, produk tersebut merupakan campuran di mana substansi lama bertahan dan bukan substansi yang baru.
c) Ketika 2 jenis molekul bereaksi, itu akan menghasilkan 1 molekul yang saling mengikat
d) Selama reaksi kimia, atom tetap sama tetapi mengatur ulang untuk membentuk molekul baru
e) ....

25. Atom diwakili oleh lingkaran, dan lingkaran berwarna yang berbeda mewakili berbagai jenis atom. Manakah dari gambar-gambar berikut ini yang merupakan senyawa kimia?

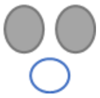
a) Gambar A



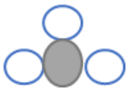
b) Gambar B



c) Gambar C



d) Gambar D



Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Senyawa kimia terdiri dari dua atom atau unsur yang terpisah satu sama lain.
b) Senyawa kimia terdiri dari dua atom atau unsur dengan tipe yang sama.
c) Senyawa kimia terdiri dari dua atom atau lebih yang berbeda dan terpisah satu sama lain.
d) Senyawa kimia terdiri dari dua atom atau lebih yang berbeda dan saling berhubungan.
e) ....

26. Manakah dari nama unsur-unsur di bawah ini yang merupakan senyawa kimia?

a) H_2 dan CO_2
b) O dan O_3
c) CH_4 dan H_2O
d) C dan H

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Senyawa kimia terdiri dari dua atom atau unsur yang terpisah satu sama lain
b) Senyawa kimia terdiri dari dua atom atau lebih yang berbeda dan saling berhubungan

c) Senyawa kimia terdiri dari dua atom atau unsur dengan tipe yang sama

d) Senyawa kimia terdiri dari dua atom atau lebih yang berbeda dan terpisah satu sama lain.

e) ....

27. In the following chemical equilibrium reaction,
2CO + O2 $\rightleftarrows$ 2CO2 $\Delta H = -40kj$

Manakah tindakan di bawah ini yang menyebabkan reaksi bergerak ke kiri?

a) Konsentrasi CO ditambahkan.

b) Temperatur menurun.

c) Volume dinaikkan.

d) Tekanan meningkat.

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Jika konsentrasi suatu zat ditambahkan, keseimbangan akan bergeser ke arah yang berlawanan.

b) Jika suhu menurun, reaksi akan bergeser ke arah reaksi eksotermik.

c) Jika volumenya naik, reaksi akan bergeser ke arah reaksi yang memiliki koefisien lebih besar.

d) Tekanan berbanding terbalik dengan volume.

e) ....

28. Dalam reaksi kesetimbangan kimia berikut,
Fe_2O_3 + 3CO $\rightleftarrows$ 2Fe + 3CO_2 $\Delta H = +30kj$

Manakah tindakan di bawah ini yang menyebabkan reaksi bergerak ke kiri?

a) Konsentrasi Fe_2O_3 ditambahkan.

b) Temperatur menurun.

c) Konsentrasi CO berkurang.

d) Katalis (Ag) ditambahkan.

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Jika konsentrasi suatu zat ditambahkan, keseimbangan akan tetap seimbang.

b) Jika suhu menurun, reaksi akan tetap seimbang.

c) Jika konsentrasi suatu zat berkurang, keseimbangan akan tetap seimbang.

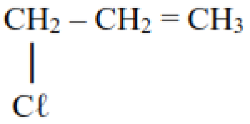
d) Katalis dapat mempercepat keseimbangan, tetapi tidak dapat menggeser kesetimbangan.

e) ....

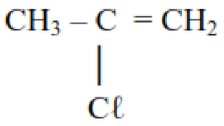
29. Manakah gambar berikut yang menunjukkan senyawa karbon yang memiliki ikatan atom karbon tersier?

a) Gambar A
$CH_3\text{- }CH = CHC\ell$

b) Gambar B



c) Gambar C

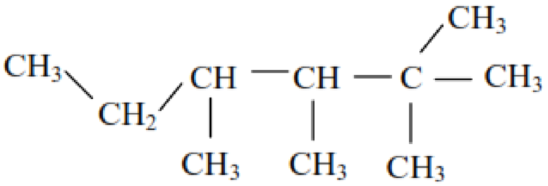


d) Gambar D
$CH_2C\ell\text{ - }CH = CH_2$

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Atom karbon tersier adalah atom karbon yang berikatan dengan atom hidrogen.
b) Atom karbon tersier adalah jumlah atom karbon yang mengikat dua atom karbon lainnya.
c) Atom karbon tersier adalah atom karbon yang mengikat tiga atom karbon lainnya
d) Atom karbon tersier adalah atom karbon yang berikatan dengan atom klorin.
e) ....

30. The following figure is a formula for the structure of carbon compounds.



Berapa jumlah atom karbon tersier dalam senyawa karbon tersebut?

a) 1
b) 2
c) 3
d) 10

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Atom karbon tersier adalah jumlah atom karbon dalam senyawa karbon.
b) Atom karbon tersier adalah jumlah atom karbon yang mengikat empat atom karbon lainnya.
c) Atom karbon tersier adalah jumlah atom karbon yang mengikat dua atom karbon lainnya.
d) Atom karbon tersier adalah atom karbon yang mengikat tiga atom karbon lainnya.
e) ....

31. Manakah senyawa kimia dengan angka oksidasi Br tertinggi?

a) $Fe(BrO_2)_3$
b) $Ca(BrO)_2$
c) $HBrO_4$
d) $AlBr_3$

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Angka oksidasi Br pada BrO_4^- = +7
b) Angka oksidasi Br pada BrO_2^- = +8
c) Angka oksidasi Br pada BrO^- = +9
d) Angka oksidasi Br pada Br^- = +7
e) ....

Apakah kamu yakin dengan jawaban yang kamu berikan pada dua pertanyaan sebelumnya?
a) Sangat yakin
b) Yakin
c) Tidak yakin
d) Sangat tidak yakin

32. Manakah reaksi kimia berikut ini yang menunjukkan reaksi redoks?

a) $AgNO_3$ + NaCl → AgCl + $NaNO_3$
b) Cl_2 + SO_2 + H_2O → HCl + H_2SO_4
c) MgO + H_2O → Cu_2 + H_2O
d) CuO + 2H → Cu_2 + H_2O

Manakah dari pernyataan berikut ini yang menjadi alasan jawaban kamu untuk pertanyaan sebelumnya?

a) Reaksi redoks adalah jumlah reaksi oksidasi dan reduksi dalam reaksi kimia.
b) Reaksi redoks adalah reaksi kimia di mana terdapat reaksi oksidasi dalam bentuk peningkatan oksidasi dan reduksi dalam bentuk penurunan jumlah oksidasi.
c) Reaksi reduksi adalah penambahan elektron dari molekul, atom, atau ion, sehingga reaksi redoks harus memiliki reaksi reduksi.
d) Reaksi oksidasi adalah pelepasan elektron dari molekul, atom, atau ion, sehingga reaksi redoks harus memiliki reaksi oksidasi.
e) ....

Appendix 5: inductive reasoning test in Indonesian version

Bagian B. Tes Kemampuan Penalaran Induktif

Selamat datang pada tes kemampuan penalaran induktif! Tujuan tes ini untuk mengetahui cara berpikir dan kemampuan penalaran Anda. Tes ini merupakan tes kemampuan Anda untuk mengaplikasikan aspek-aspek penalaran induktif, menganalisis situasi dan membuat prediksi atau untuk menyelesaikan suatu masalah.

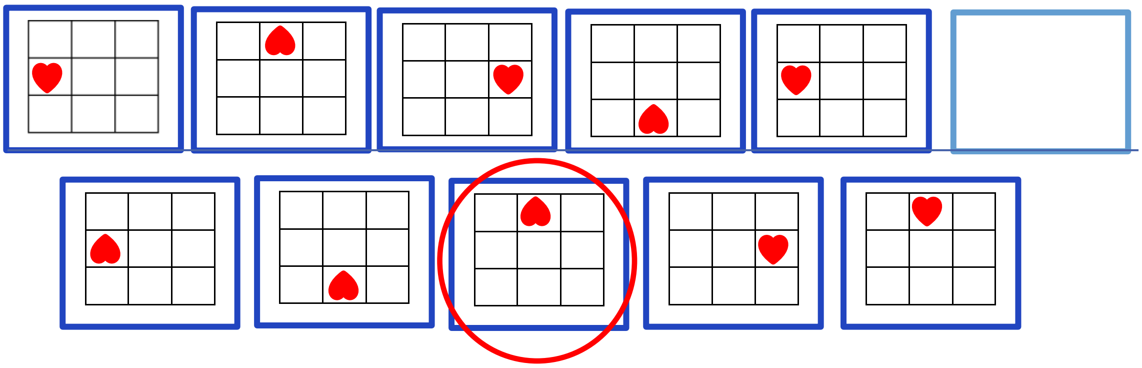
PETUNJUK

- Soal ini terdiri dari 4 bagian, pada setiap bagian soal tersedia contoh dan tips untuk menjawab soal.
- Baca petunjuk pengerjaan soal dengan hati-hati sebelum mengerjakan soal. Semoga sukses!

Bagian A.

Temukan peraturan untuk gambar berikutnya.

Contoh: Pilihlah gambar yang paling tepat untuk bingkai kuning.



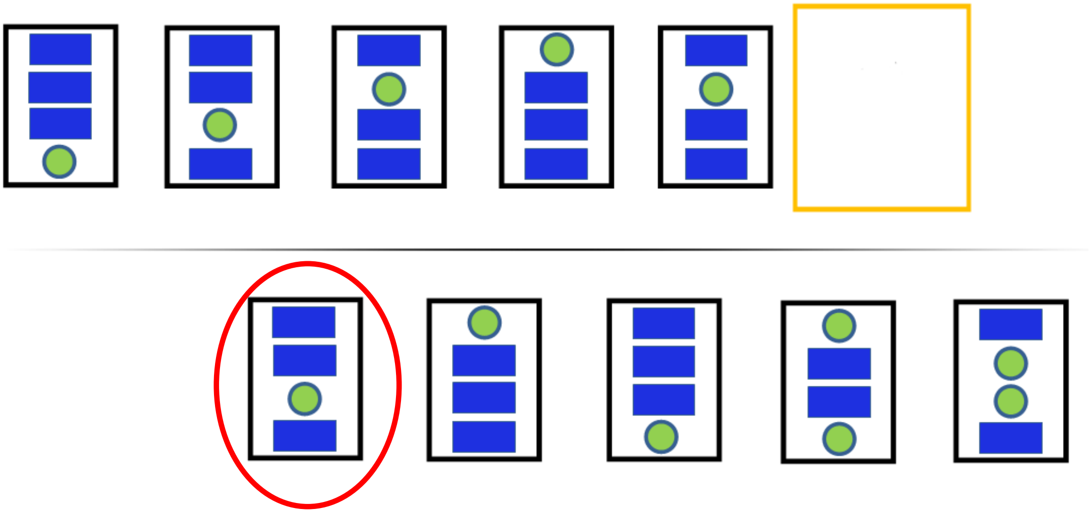
Tips untuk menjawab:

Peraturan Pertama: Gambar berbentuk hati bergerak 2 kotak setiap waktu searah jarum jam.

Peraturan Kedua: Arah gambar bebrbentuk hati berubah secara berlawanan pada setiap bingkai berikutnya. Jadi, jawaban yang benar adalah gambar ketiga.

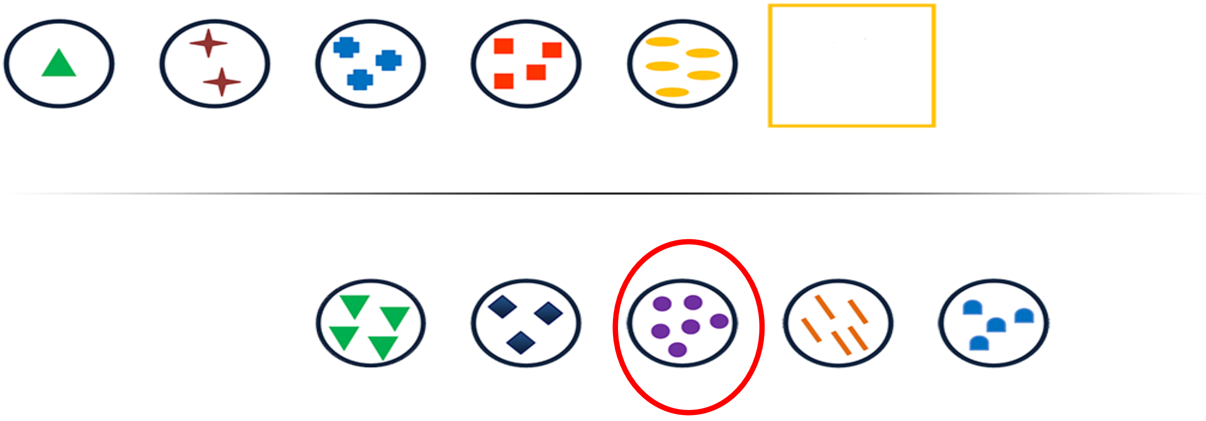
Soal No. 1 (IR_FS01)

Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.



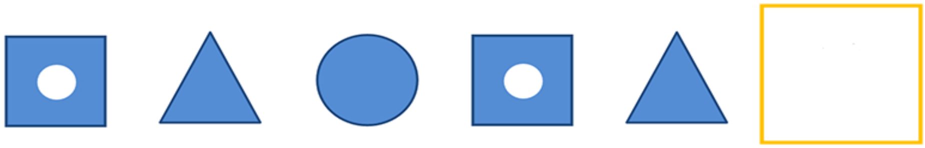
Soal No. 2 (IR_FS02)

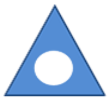
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

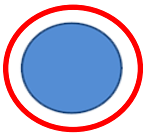


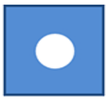
Soal No. 3 (IR_FS03)

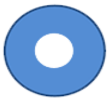
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

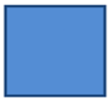






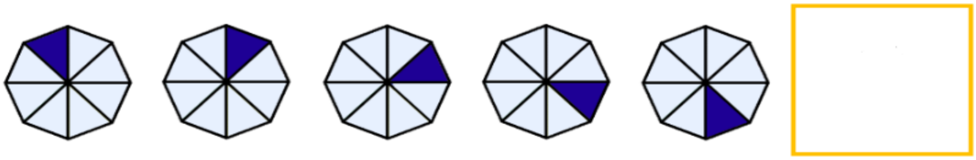


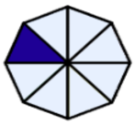


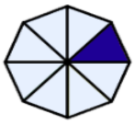


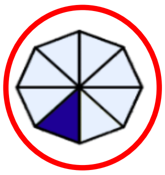
Soal No. 4 (IR_FS04)

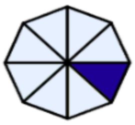
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

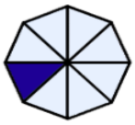






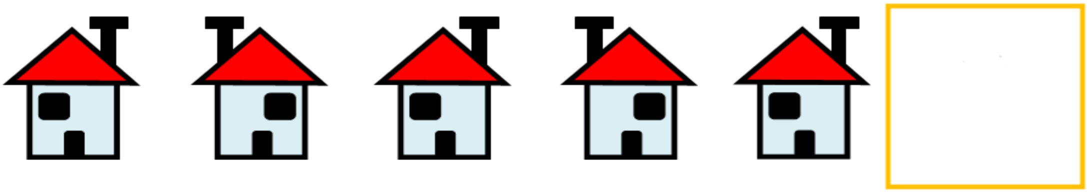


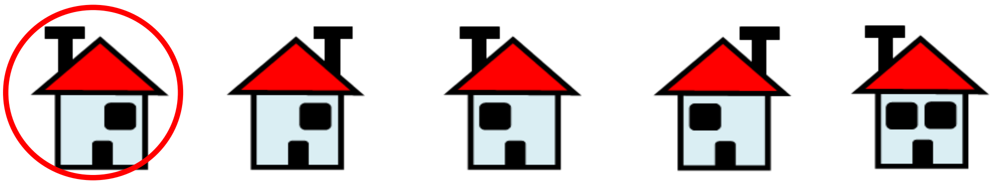




Soal No. 5 (IR_FS05)

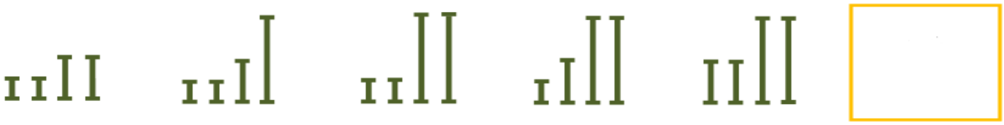
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

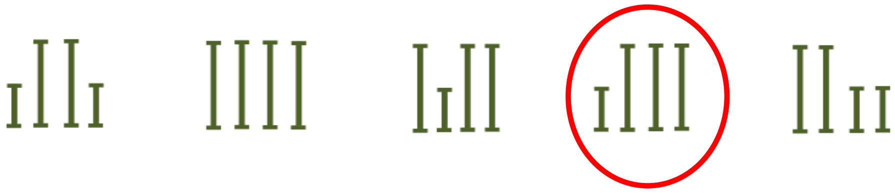




Soal No. 6 (IR_FS06)

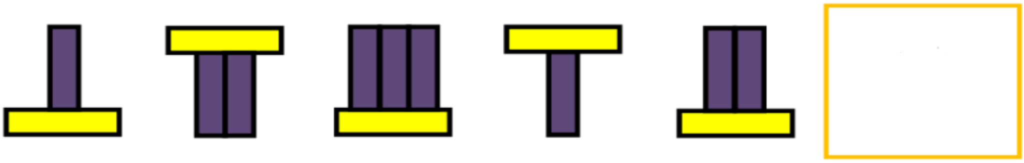
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

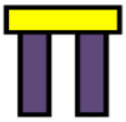


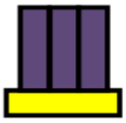


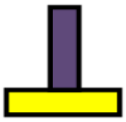
Soal No. 7 (IR_FS07)

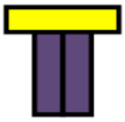
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

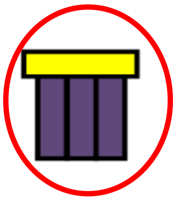






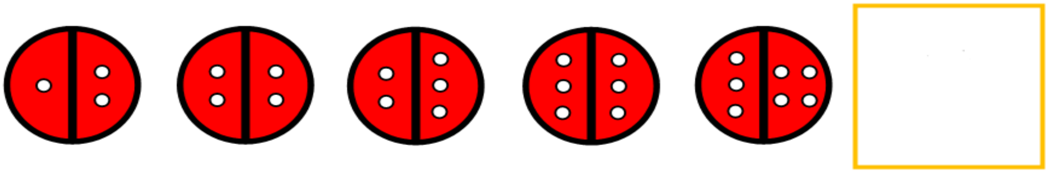


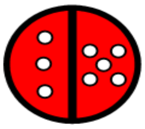


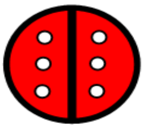


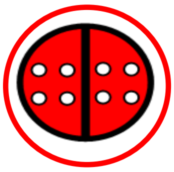
Soal No. 8 (IR_FS08)

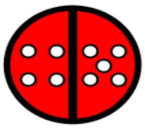
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

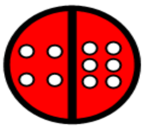






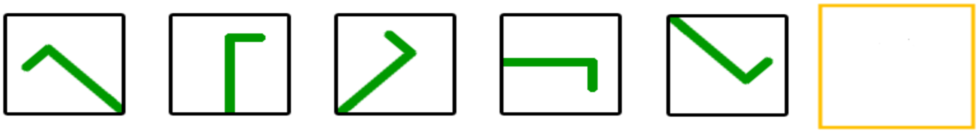


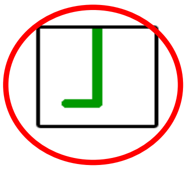


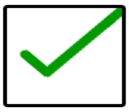


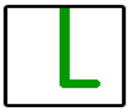
Soal No. 9 (IR_FS09)

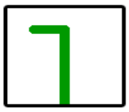
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

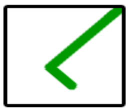






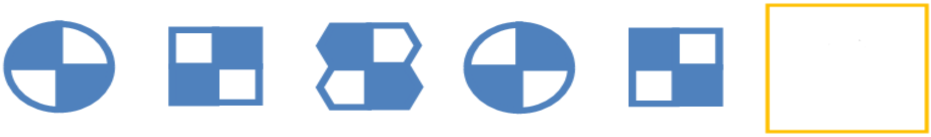


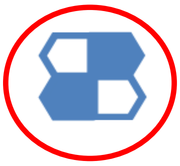


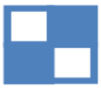


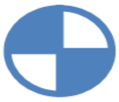
Soal No. 10 (IR_FS10)

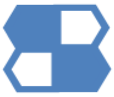
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

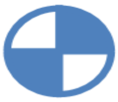








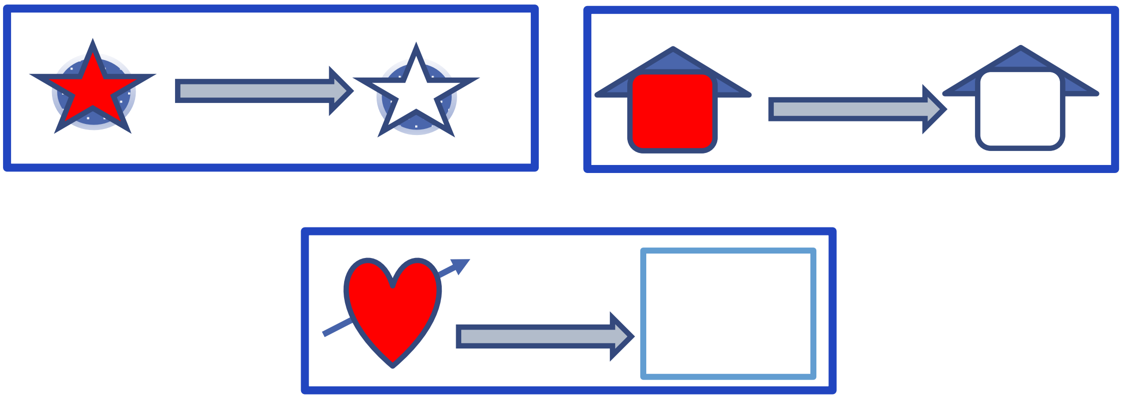


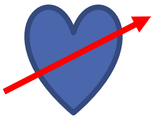


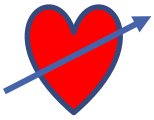
Bagian B:

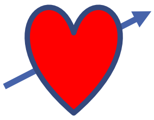
Pada bagian ini, periksalah bagaimana angka berubah pada bingkai biru! Bagaimana aturannya?

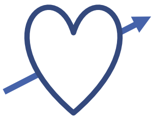
Contoh: Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

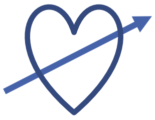












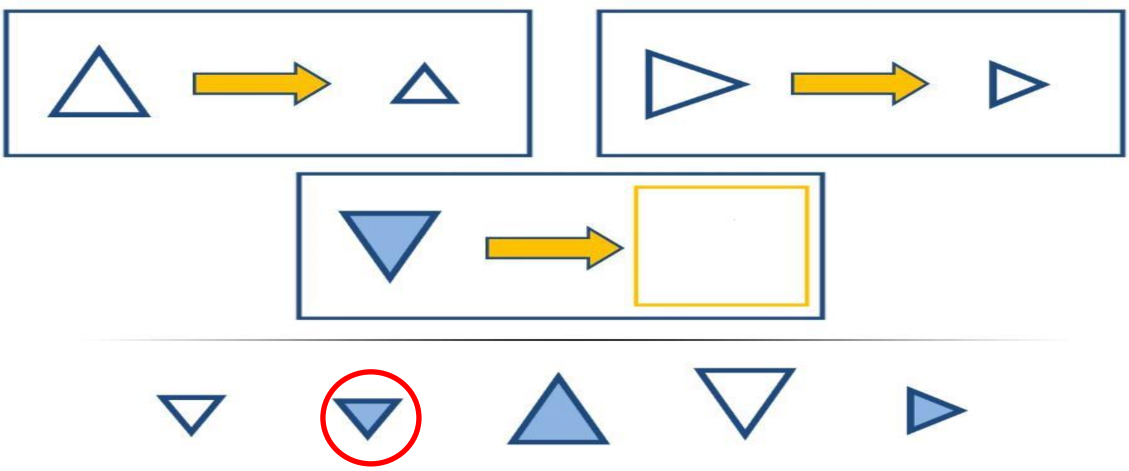
Tips untuk menjawab:

Aturan pertama: Dua gambar memiliki warna yang berbeda (biru dan merah), tetapi hanya gambar merah yang berubah menjadi berwarna putih.

Aturan kedua: The blue shape is always in the back of other shape. Jadi, jawaban yang benar adalah gambar keempat.

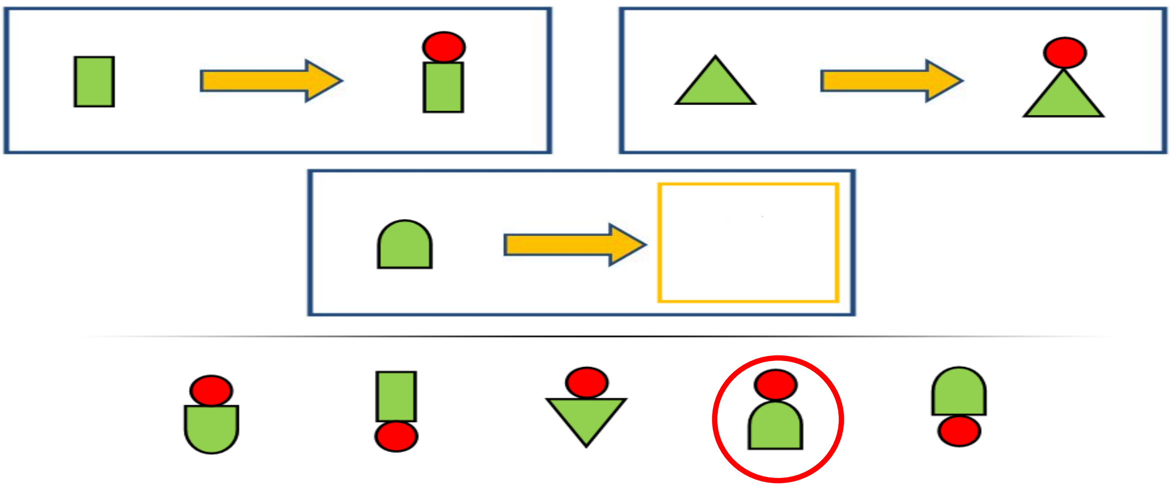
Soal No. 11 (IR_FA01)

Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.



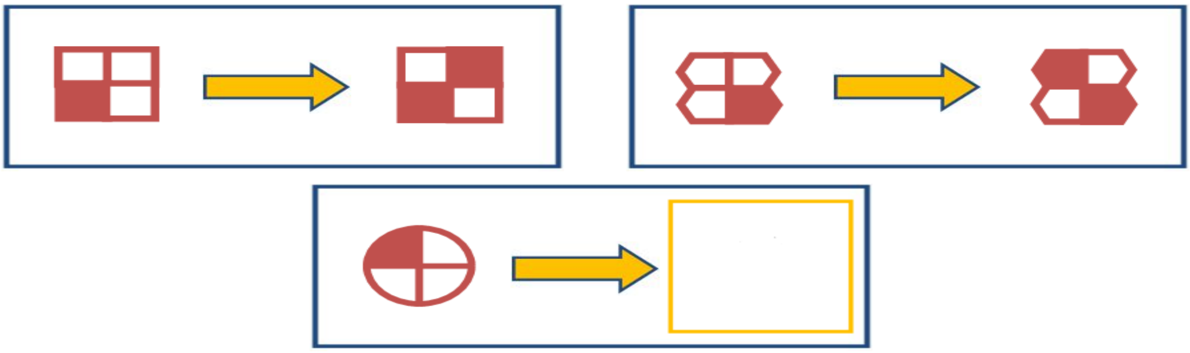
Soal No. 12 (IR_FA02)

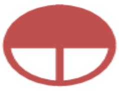
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

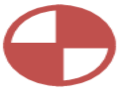


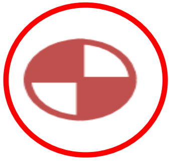
Soal No. 13 (IR_FA03)

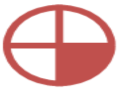
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

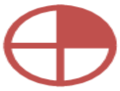






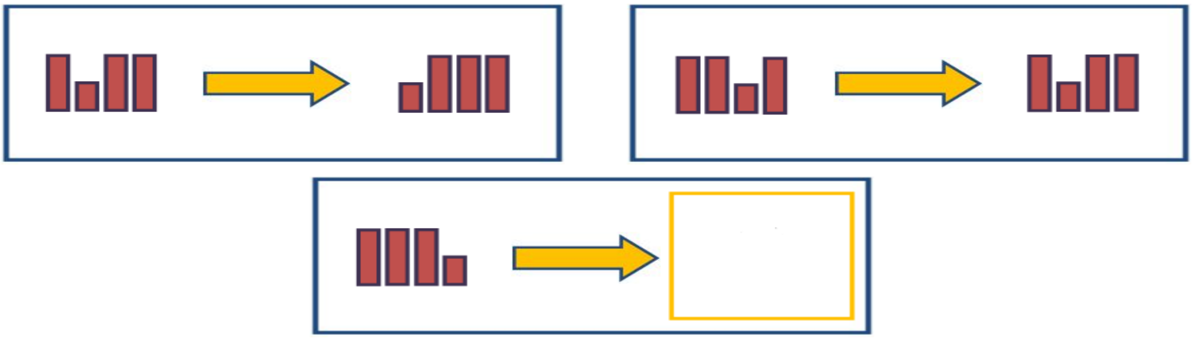


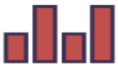


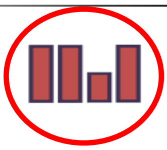


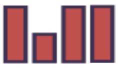
Soal No. 14 (IR_FA04)

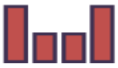
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

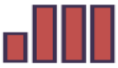






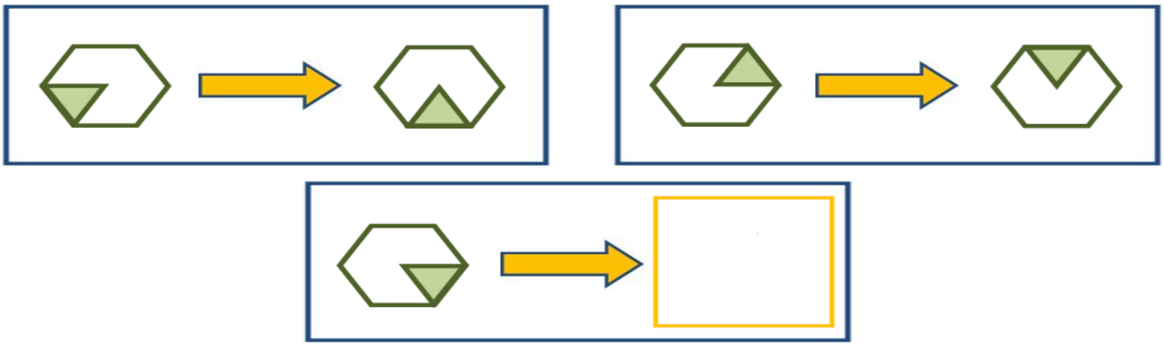


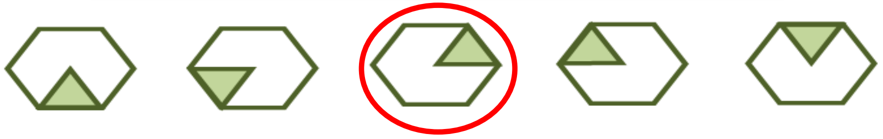




Soal No. 15 (IR_FA05)

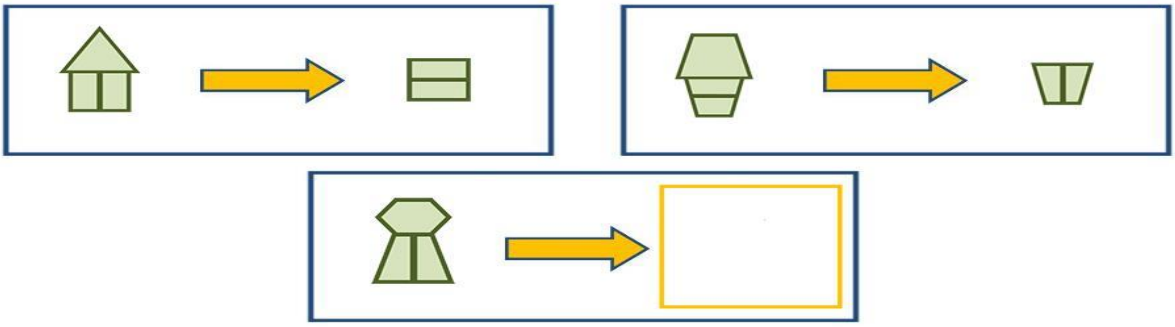
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

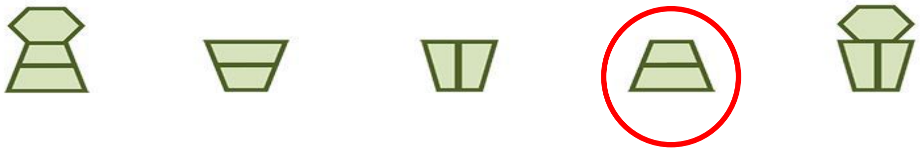




Soal No. 16 (IR_FA06)

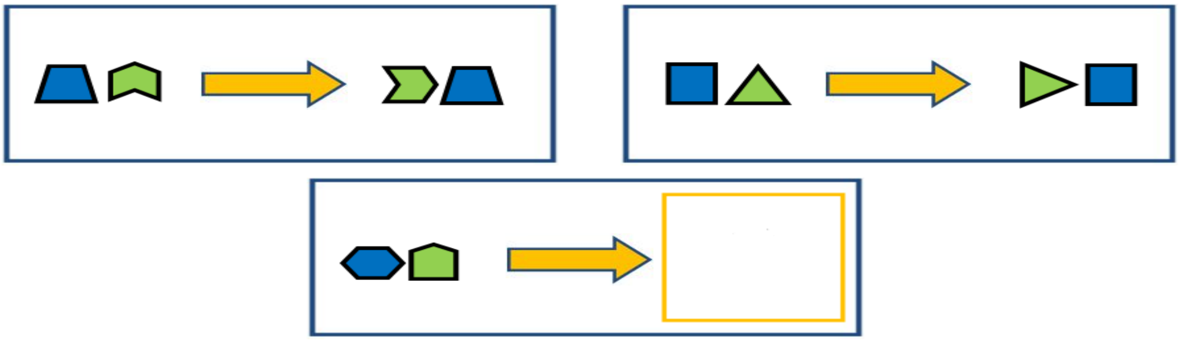
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

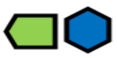


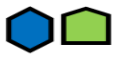


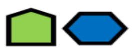
Soal No. 17 (IR_FA07)

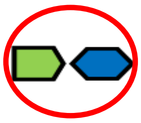
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

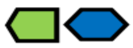






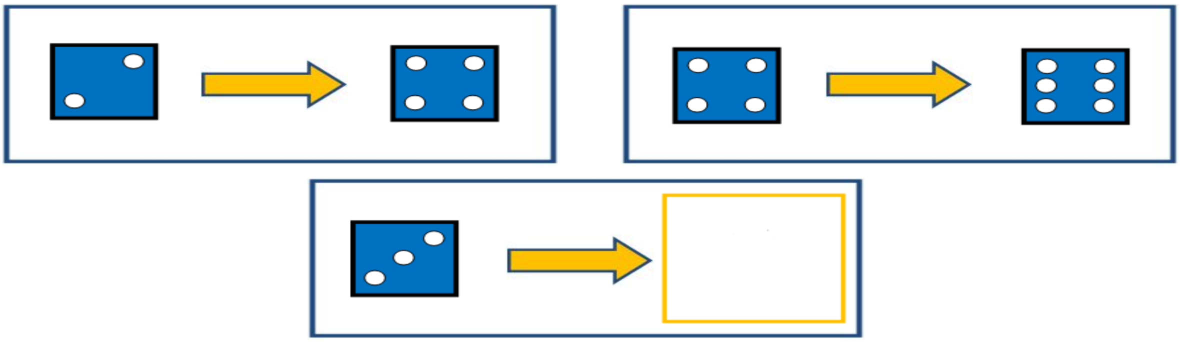


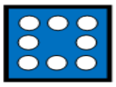


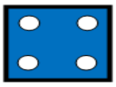


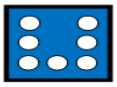
Soal No. 18 (IR_FA08)

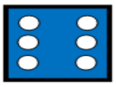
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

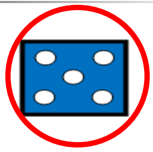






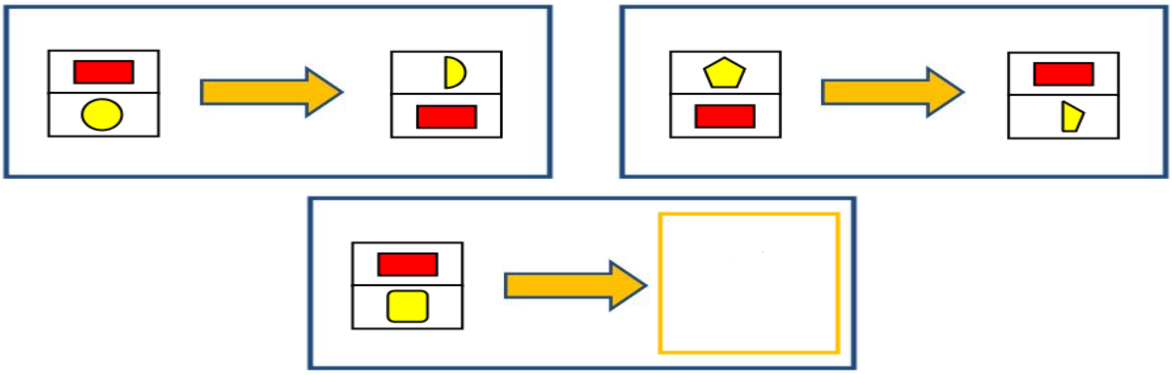


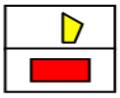


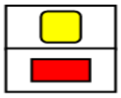


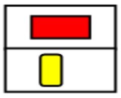
Soal No. 19 (IR_FA09)

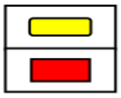
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

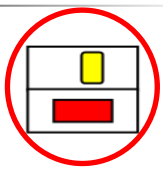






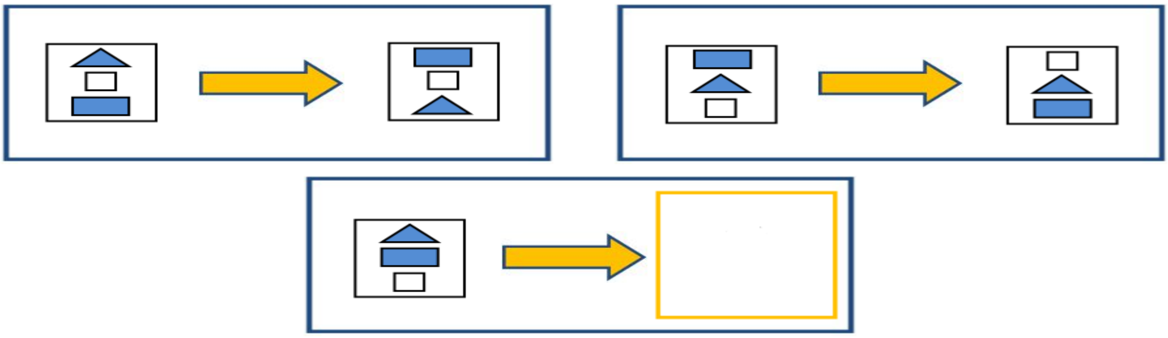


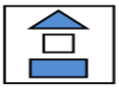


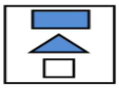


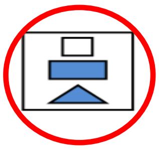
Soal No. 20 (IR_FA10)

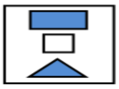
Gambar manakah yang paling cocok pada bingkai kuning? Pilih dan geser gambar tersebut ke bingkai kuning.

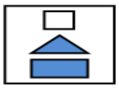






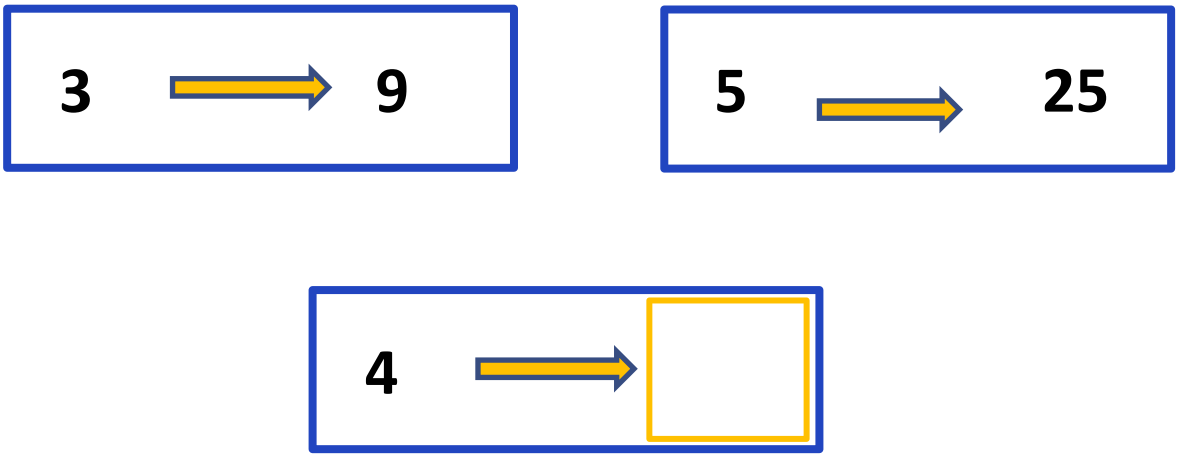






Bagian C: Pada bagian ini, periksalah bagaimana angka berubah pada bingkai biru! Bagaimana aturannya?

Contoh: Angka manakah yang paling cocok pada bingkai kuning? Pilih dan geser angka tersebut ke bingkai kuning.



10 16 25 20 30

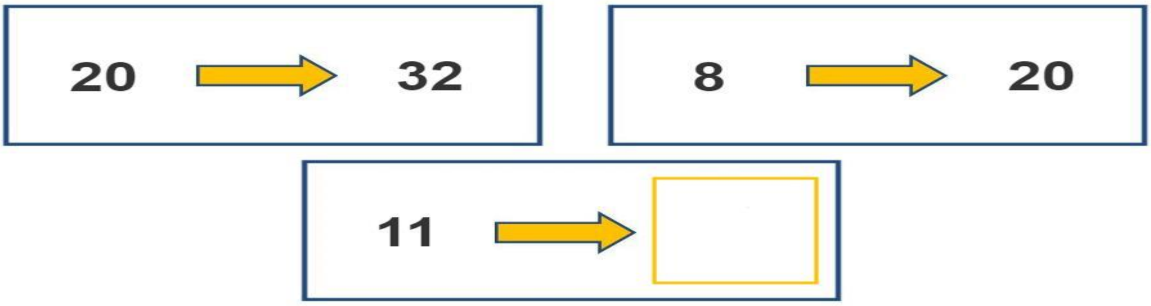
Tips untuk menjawab:

Peraturan: Pada bingkai biru, angka yang kedua sama dengan kuadrat dari angka pertama.

Jadi, jawaban yang benar adalah pilihan yang kedua (16).

Soal No. 21 (IR_NA01)

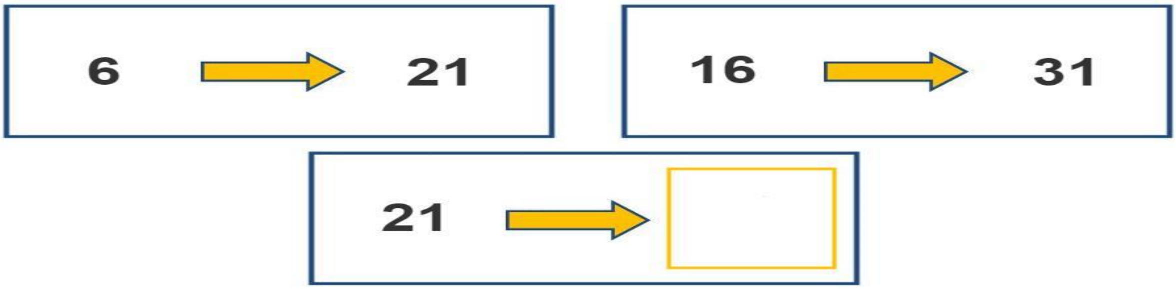
Angka manakah yang paling cocok pada bingkai kuning? Pilih dan geser angka tersebut ke bingkai kuning..



20 [✓ 23] 22 33 12

Soal No. 22 (IR_NA02)

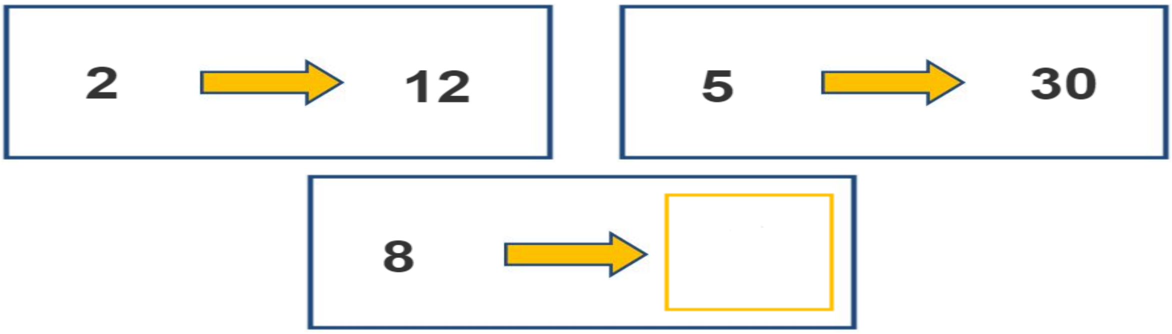
Angka manakah yang paling cocok pada bingkai kuning? Pilih dan geser angka tersebut ke bingkai kuning.



31 41 6 15 [✓ 36]

Soal No. 23 (IR_NA03)

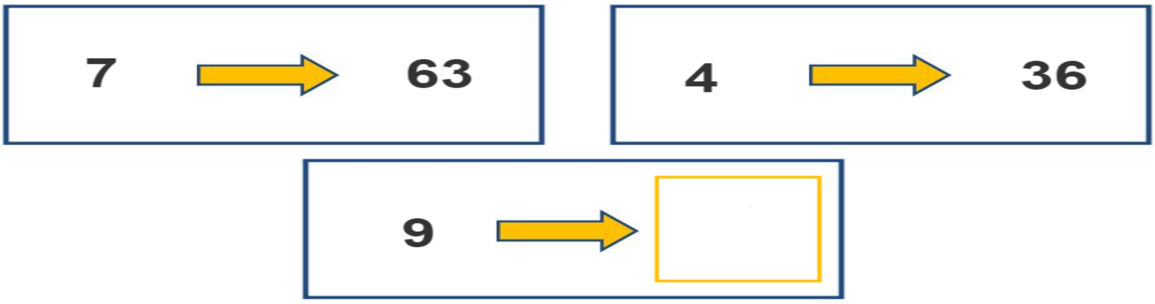
Angka manakah yang paling cocok pada bingkai kuning? Pilih dan geser angka tersebut ke bingkai kuning.



28 18 [✓ 48] 40 38

Soal No. 24 (IR_NA04)

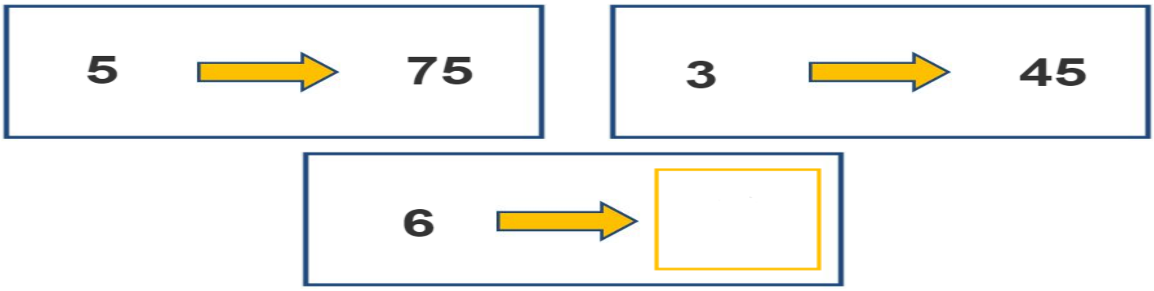
Angka manakah yang paling cocok pada bingkai kuning? Pilih dan geser angka tersebut ke bingkai kuning.



65 [✓ 81] 41 36 63

Soal No. 25 (IR_NA05)

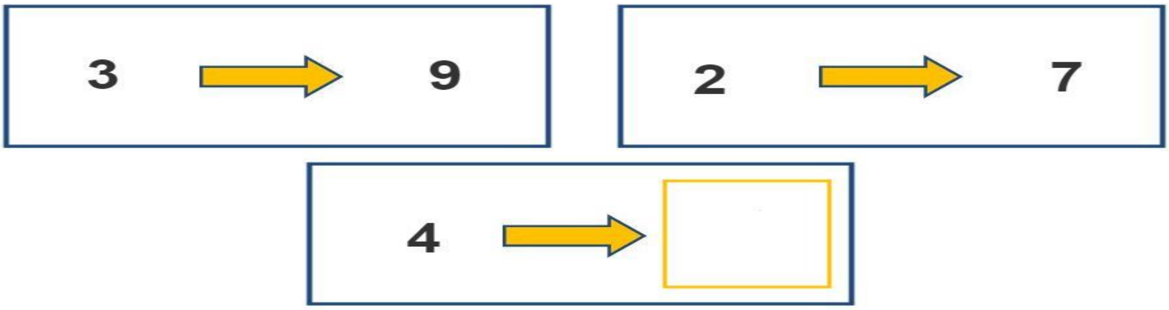
Angka manakah yang paling cocok pada bingkai kuning? Pilih dan geser angka tersebut ke bingkai kuning.



[✓ 90]   76   80   78   30

Soal No. 26 (IR_NA06)

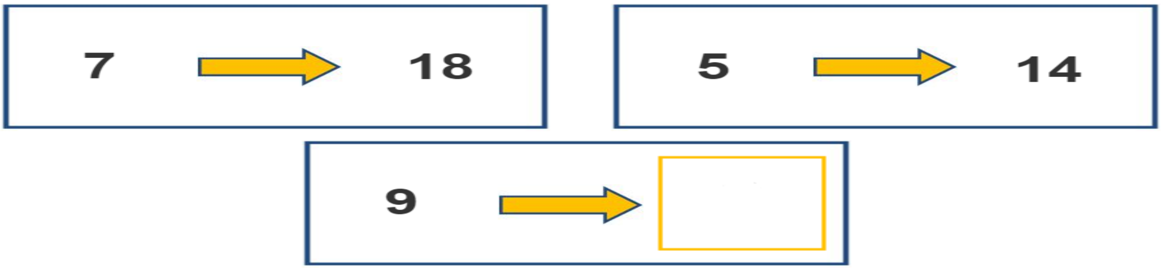
Angka manakah yang paling cocok pada bingkai kuning? Pilih dan geser angka tersebut ke bingkai kuning.



10   14   12   [✓ 11]   8

Soal No. 27 (IR_NA07)

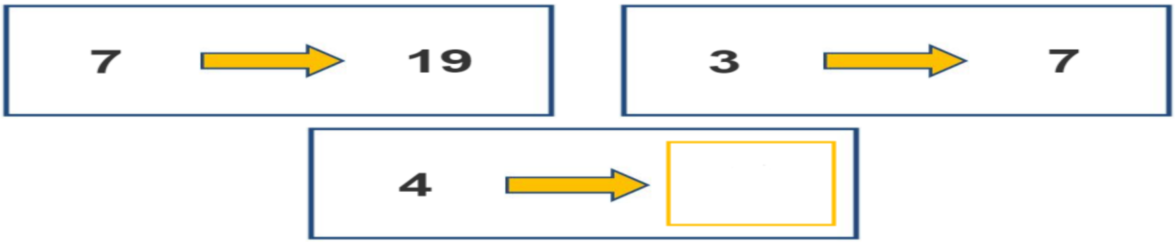
Angka manakah yang paling cocok pada bingkai kuning? Pilih dan geser angka tersebut ke bingkai kuning.



16 [✓ 22] 20 18 10

Soal No. 28 (IR_NA08)

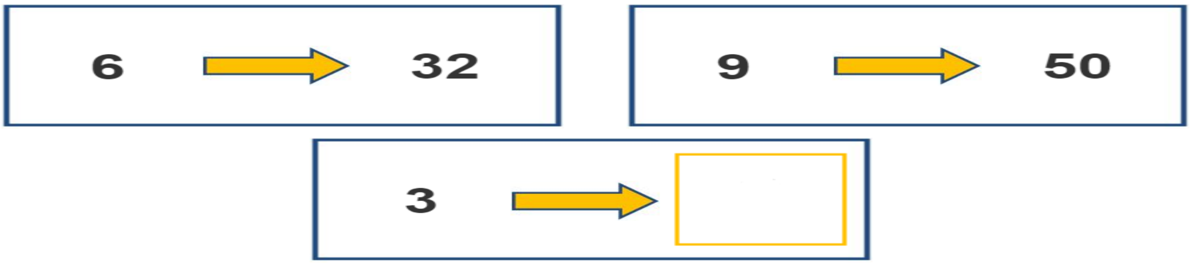
Angka manakah yang paling cocok pada bingkai kuning? Pilih dan geser angka tersebut ke bingkai kuning.



9 8 [✓ 10] 12 16

Soal No. 29 (IR_NA09)

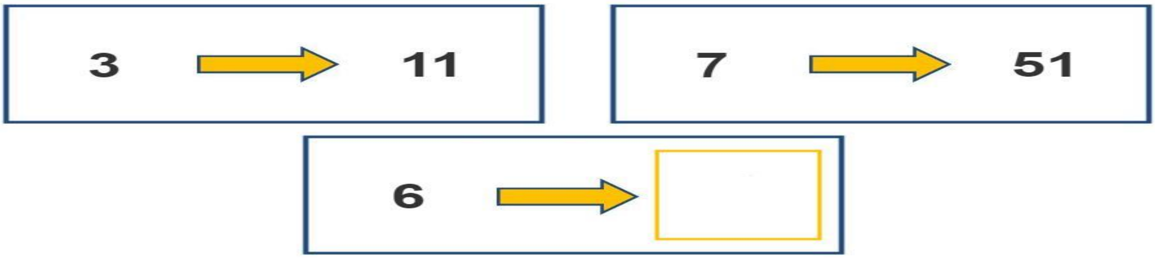
Angka manakah yang paling cocok pada bingkai kuning? Pilih dan geser angka tersebut ke bingkai kuning.



18 [✓ 14] 29 16 23

Question 30 (IR_NA10)

Angka manakah yang paling cocok pada bingkai kuning? Pilih dan geser angka tersebut ke bingkai kuning.



41 14 42 91 [✓ 38]

Bagian D:

Pada bagian ini, diperlukan dua angka untuk melanjutkan deret bilangan!

Contoh: Dua angka harus dimasukkan ke dalam bingkai kuning untuk menyelesaikan deret bilangan berikut! Pilih dan seret angka ke dalam bingkai kuning.

1	**3**	**5**	**7**	**9**	**11**		

27 [✓ **13**] [✓ **15**] **21** **25** **19** **17**

Tips untuk menjawab:

Peraturan: Ini merupakan deret bilangan ganjil.

Jadi, jawaban yang benar berturut-turut adalah 13 dan 15.

Soal No. 31 (IR_NS01)

Dua angka harus dimasukkan ke dalam bingkai kuning untuk menyelesaikan deret bilangan berikut! Pilih dan seret angka ke dalam bingkai kuning.

1	2	4	8	16	32		

33 **48** **40** [✓ **64**] **35** [✓ **128**] **124**

Soal No. 32 (IR_NS02)

Dua angka harus dimasukkan ke dalam bingkai kuning untuk menyelesaikan deret bilangan berikut! Pilih dan seret angka ke dalam bingkai kuning.

0	1	3	4	6	7		

[✓ 10] 13 8 11 [✓ 9] 12 14

Soal No. 33 (IR_NS03)

Dua angka harus dimasukkan ke dalam bingkai kuning untuk menyelesaikan deret bilangan berikut! Pilih dan seret angka ke dalam bingkai kuning.

3	6	11	14	19	22		

26 28 [✓ 30] 25 32 [✓ 27] 29

Soal No. 34 (IR_NS04)

Dua angka harus dimasukkan ke dalam bingkai kuning untuk menyelesaikan deret bilangan berikut! Pilih dan seret angka ke dalam bingkai kuning.

1	2	4	7	11	16		

26 17 20 19 [✓ 22] 21 [✓ 29]

Soal No. 35 (IR_NS05)

Dua angka harus dimasukkan ke dalam bingkai kuning untuk menyelesaikan deret bilangan berikut! Pilih dan seret angka ke dalam bingkai kuning.

1	2	3	5	8	13		

14 [✓ 21] 18 [✓ 34] 19 23 26

Soal No. 36 (IR_NS06)

Dua angka harus dimasukkan ke dalam bingkai kuning untuk menyelesaikan deret bilangan berikut! Pilih dan seret angka ke dalam bingkai kuning.

3	9	7	15	11	21		

16 25 31 41 [✓ 27] 19 [✓ 15]

Soal No. 37 (IR_NS07)

Dua angka harus dimasukkan ke dalam bingkai kuning untuk menyelesaikan deret bilangan berikut! Pilih dan seret angka ke dalam bingkai kuning.

3	5	9	17	33	65		

99 61 109 [✓ 257] [✓ 129] 57 217

Soal No. 38 (IR_NS08)

Dua angka harus dimasukkan ke dalam bingkai kuning untuk menyelesaikan deret bilangan berikut! Pilih dan seret angka ke dalam bingkai kuning.

1	10	26	51	87	136		

145 161 146 162 151 [✓ 200] [✓ 281]

Soal No. 39 (IR_NS09)

Dua angka harus dimasukkan ke dalam bingkai kuning untuk menyelesaikan deret bilangan berikut! Pilih dan seret angka ke dalam bingkai kuning.

1	3	10	24	47	81		

109 84 [✓ 128] 94 83 90 [✓ 190]

Soal No. 40 (IR_NS10)

Dua angka harus dimasukkan ke dalam bingkai kuning untuk menyelesaikan deret bilangan berikut! Pilih dan seret angka ke dalam bingkai kuning.

1	2	4	8	15	26		

29 [✓ 42] [✓ 64] 37 48 27 45

Terima kasih atas partisipasinya

- **Selesai -**

RELEVANT PUBLICATIONS

Related publications in dissertation

Soeharto, S. (2021). Evaluation and development of students' misconception using diagnostic assessment in science across school grades: A Rasch measurement approach. Journal of Turkish Science Education, 18(3), 351-370. https://doi.org/10.36681/tused.2021.78

Soeharto, S., & Csapó, B. (2021). Evaluating item difficulty patterns for assessing student misconceptions in science across physics, chemistry, and biology concepts. Heliyon, 7(11), e08352. https://doi.org/10.1016/j.heliyon.2021.e08352

Soeharto, S., & Csapó, B. (2022a). Exploring Indonesian student misconceptions in science concepts. Heliyon, 8(9), e10720. https://doi.org/10.1016/j.heliyon.2022.e10720

Soeharto, S., & Csapó, B. (2022b). Assessing Indonesian student inductive reasoning: Rasch analysis. Thinking Skills and Creativity, 46, 101132. https://doi.org/10.1016/j.tsc.2022.101132

Soeharto, S., Csapő, B., Sarimanah, E., Dewi, F. I., & Sabri, T. (2019). A Review of Students' Common Misconceptions in Science and Their Diagnostic Assessment Tools. Jurnal Pendidikan IPA Indonesia, 8(2). https://doi.org/10.15294/jpii.v8i2.18649

Conference papers

Soeharto, S., & Csapó, B. (2019). Students' Misconceptions and Diagnostic Assessment in Science. In: Varga, Aranka; Andl, Helga; Molnár-Kovács, Zsófia (eds.) Neveléstudomány – Horizontok és dialógusok. Absztraktkötet. : XIX. ONK. Pécs, 2019. november 7-9. Pécs, Hungary : Pécsi Tudományegyetem Bölcsészettudományi Kar Neveléstudományi Intézet, pp. 538-538.

Soeharto, S. (2019). Developing Three-Tier Diagnostic Test In Science To Assess Student Misconceptions In Science. Abstract book: The 13th Training and Practice International Conference on Educational Science. Kaposvár, Hungary: Kaposvár University Faculty of Pedagogy, pp. 202-202.

Soeharto, S. (2021). The Evaluation of Students` Inductive Reasoning and Its Role In Science Achievement Using Rasch Analysis. Abstract book: The 14th Training

and Practice International Conference on Educational Science. Kaposvár, Hungary: Kaposvár University Faculty of Pedagogy, pp. 55-55.

Soeharto, S., & Csapó, B. (2021). Psychometric Evaluation in Developing E-Learning Readiness in Science Classroom (ELRSC) Questionnaire Using Rasch Analysis. Abstract book: ATEE-EDITE-ELTE online conference on 11 June: Research in Teacher Education – the next generation. Hungary: Budapest, 48-48.

Soeharto, S., & Csapó, B. (2021). Investigating the relationship between test anxiety and motivation in science learning. Abstract book: in JURE 2021: Education and Citizenship: Learning and Instruction and the Shaping of Futures. Online, pp. 24-24.

Soeharto, S., & Csapó, B. (2021). The diagnostic test evaluation and the student misconception development in science. Abstract book: in EARLI 2021: Education and Citizenship: Learning and Instruction and the Shaping of Futures. Online, pp. 217-217.

Soeharto, S., & Csapó, B. (2021). Investigating students' e-learning readiness in the science classroom. Abstract book: in EAPRIL 2021: the 15th annual EAPRIL Conference for Practitioner Research on Improving Learning. Online, pp. 42-42.

Soeharto, S. (2021). *Evaluating Students' E-Learning Training Needs in Science Classroom during COVID-19 Pandemic Era*. In: Molnár, Győnyvér and Tóth Edith; Dancs, Katinka (eds.) *CES 2021: 21st Conference on Educational Sciences: Education's responses to future challenges*. Programme and Abstracts. Szeged, Hungary: Szegedi Tudományegyetem, pp. 477-477.

Soeharto, S. (2021). *Investigating the Influence of Students' Inductive Reasoning on Science and Mathematics Achievement.* In: Molnár, Győnyvér and Tóth Edith; Dancs, Katinka (eds.) *CES 2021: 21st Conference on Educational Sciences: Education's responses to future challenges*. Programme and Abstracts. Szeged, Hungary: Szegedi Tudományegyetem, pp. 287-287.

DECLARATION

I certify that the dissertation's substance is entirely my own original works from published systematic review and empirical studies. This dissertation has never been presented to another university or for a different degree.